

The International Journal of Science | 10 October 2019

outlook  
Prostate  
cancer

# nature

## ROOTS OF DIVERSITY

Transcriptome analysis illuminates evolution of the world's green plants

**A history of ethics**  
The long and bumpy road to responsible research

**Ancient climate**  
A snapshot of CO<sub>2</sub> in the atmosphere more than 1 million years ago

**Species in decline**  
Ten-year surveys reflect strong evidence of falling numbers



## Moedas's legacy — and what Europe must do next

**The successor to the European Union's research chief must act to prevent budget cuts.**

**C**arlos Moedas was little known outside Portugal when he took over as the European Union's research and innovation chief in 2014.

Now, at the end of his tenure, that is no longer the case. In five years, the engineer-turned-banker-turned politician has demonstrated thoughtful advocacy for research. He has listened to researchers and delivered — except on one issue where it really matters.

The funding settlement for Horizon Europe, the next research framework programme for all EU member states, has hit a roadblock. Moedas's successor, Mariya Gabriel, and Europe as a whole must work hard to fight cuts and potential delays to its start.

On the positive side of the ledger, it is because of Moedas that around €9 billion (US\$10 billion) — around one-tenth of the next round of European research funding — will be set aside for large collaborations in five global challenges or 'missions' — in climate change, cancer, oceans, smart cities, and soil and food. This was an idea that Moedas adopted after discussions with researchers, notably the innovation economist Mariana Mazzucato.

But a European research commissioner's core job — some would argue the most important one — is to protect the budget. Earlier this month, negotiations between EU member states on the next seven-year budget cycle (for 2021–27) stalled. The European Commission is asking for €1.135 trillion, including around €100 billion for research. Member states want to cut the total budget by between €35 billion and €85 billion. Facing such a shortfall, it isn't uncommon for those in charge of setting budgets to look to research for cuts.

Protecting research needs firepower — it requires support from heads of government, and especially from national ministries of finance. Moedas and his boss, commission president Jean-Claude Juncker, should have assembled high-level support much earlier, before we got to this point. The responsibility for ensuring that research does not bear the brunt of any cuts now falls to Gabriel.

An added complication is that, under the incoming commission, the department for research and innovation is being merged with that for education, youth, sport and culture. This expanded department is called Innovation and Youth — 'research' has been lost from the title — and Gabriel will have extra, and possibly competing, priorities, one of which is a trebling of the budget for the student-exchange programme Erasmus+.

**“Protecting research needs firepower — it requires support from heads of government.”**

Moedas has been a popular commissioner, known as a team player and a conciliator — playing the 'good cop' to his former head of research Robert-Jan Smits's 'bad cop' in budget discussions. Smits describes Moedas as “a genuine, nice person who doesn't like to put people in an uncomfortable situation”. These are important qualities.

But the EU faces some significant challenges, and Gabriel will need to adopt a tougher persona. Economies are slowing; austerity has been painful and many governments want to spend more at home on social programmes. At the same time, budget planners will need to adjust for the potential absence of — or reduction in — the UK contribution to the EU.

If they want to see their EU research budgets protected, research organizations can help Gabriel by putting pressure on their national governments, especially finance ministries. Everyone needs to push harder to protect funding — so that the spirit and support that has helped make the EU a model for collaborative research can live on.

## Brexit promises are premature

**Government offers of new funds for UK scientists could be unaffordable.**

**T**here's a research group in Britain that has become a staple of the country's news shows, and it's called The UK in a Changing Europe. On most nights, the team of political scientists, economists and lawyers dispassionately responds to broadcasters' questions on the impact — economic, political and societal — of the United Kingdom's departure from the European Union.

The researchers, who are funded by the UK government's Economic and Social Research Council — but whose work is independent of the government's own policies — do not have an easy task. But it's an important one, in part because the government has not yet released its own detailed analysis of Brexit's impacts.

Lawmakers know that most researchers would like nothing more than for the United Kingdom to remain a member of the EU. That is one reason that The UK in a Changing Europe team, which is one of just a handful of independent analysts, is careful not to dwell on the impact of Brexit on the research community — but instead is keeping the focus on the bigger picture.

As this Editorial went to press, the EU had agreed to a request from the UK government to delay Brexit to 31 January 2020 — three months beyond the recent, 31 October, deadline. And with Prime Minister Boris Johnson and members of the Parliament at loggerheads over the terms of the exit, politicians were preparing for a general election. Researchers will have breathed a sigh of relief at avoiding an

October 'no-deal', but few will be rejoicing. In what the EU calls a "flexextension", Brexit could happen before 31 January if Parliament approves a deal.

And if a deal such as the one Johnson and the EU have agreed is ultimately passed, the worst case is that Britain leaves the free-trade area known as the customs union. Free movement of citizens to and from the EU and Britain will end, and Britain's researchers might no longer be able to obtain funding from certain EU research programmes.

That is the scenario policymakers are planning for. But as this journal – along with organizations representing researchers, such as the Royal Society – has repeatedly said, fracturing more than four decades of joint working between the United Kingdom and its nearest neighbours will damage both science and society.

Aware of these concerns – and especially of the need to maintain scientific connections – the Johnson government has been talking up post-Brexit wins for research.

It plans a more favourable visa regime for students and researchers, and is shaping new funds, including a UK version of the United States' Defense Advanced Research Projects Agency. There's also talk of a generous European Research Council-style fund for UK researchers, should access to the EU scheme no longer be possible, and more funds to collaborate with the United States. And there's confidence among some policymakers that the world's researchers will continue to want to work with, and in, Britain.

Such confidence is premature. A more welcoming visa regime and extra funding will help to placate some of researchers' concerns, but new cash depends on how much the UK Treasury department has to spend, and that relies on two things that the country does not control. The first is how much Britain will have to pay the EU for any future relationship. The second – and more important – factor is that any funding increase for research needs the UK economy to continue to grow. Although the Treasury has carried out detailed economic-impact analyses of future growth, the chancellor of the exchequer, Sajid Javid, is not yet releasing the results.

But thanks to modelling from UK in a Changing Europe, we know that, under Johnson's proposals, income per capita is projected to be 2.5% lower on average than if Britain remained an EU member, based on economists' projections of income from trade and reduced immigration. The team also says that when trade falls, which it will in the initial period after Brexit, that also reduces productivity. After factoring productivity losses into the models, post-Brexit income per capita could be between 2.3% and 7% lower.

These figures call into question assumptions that Brexit will bring an economic dividend. And without such a dividend, the government will probably struggle to keep its promises of increased research funding.

As the Brexit saga rolls on, researchers need to continue their objective analyses of its potential impacts, and to call out what could be prematurely optimistic promises. They must highlight the risks to research and ensure that none of these issues is trampled on in the stampede to get a deal in place.

“  
The slow  
growth of  
India's tiger  
population  
is a rare good  
news story.”

## Open data could save more tigers

**India has a duty to give researchers access to the raw data on this threatened species.**

**O**n Global Tiger Day in July, India's government announced a victory. It declared that the nation is home to 2,967 wild tigers – a major increase on the 1,872 animals recorded in 1972.

Centuries ago, tens of thousands of tigers roamed the world. Today, only six sub-species remain and the International Union for Conservation of Nature estimates that there are no more than 3,159 individuals in the wild. But, after centuries of hunting and habitat destruction, India's tigers seem to be turning a corner. However, as we report on page 612, there's much more that could be done.

To begin with, tiger-conservation work must be improved in more of India's 50 tiger reserves – the current effort is concentrated in just a handful. And the government must give scientists at India's universities access to the reserves and the raw data on which its tiger estimates are based.

At present, scientists cannot access the full data that the government collects during the national tiger census, which is conducted every four years. This is in spite of the fact that India's National Data Sharing and Accessibility Policy states that scientific data collected using “public funds” should be made available to those with a legitimate research interest.

If researchers were allowed to see the raw numbers, they could independently verify tiger population estimates. Such verification is essential to knowing whether conservation measures are working. It could also allow scientists to project population trends over time, and estimate birth and death rates. These measures would help officials in the government forestry department to assess populations more accurately and act quickly if they foresee a risk of local extinction. Such action might have helped to prevent the complete loss of tigers from three reserves that the 2018 census reports.

Researchers are also keen to get involved in, and improve, the census itself. The four-yearly survey is a gargantuan effort. The 2018 one covered 381,400 square kilometres and amounted to nearly 594,000 human-days of work. It logged 35 million photographs taken with hidden motion-triggered cameras, yielding almost 77,000 images of tigers.

But instead of trying to count tigers across such a vast area, individual reserves – where 70–85% of India's tigers are thought to be found – could be sampled more often, and automatic image recognition used to process the pictures.

The slow growth of India's tiger population is a rare good news story in international conservation. But the Indian government could be doing more. It must trust independent scientists with the raw data, so that one of Earth's most iconic species can survive long into the future.

# World view

## Contract cheating will erode trust in science



By Tracey Bretag

**To combat academic dishonesty, focus on educational systems and not just individual offenders, says Tracey Bretag.**

**S**tories of students paying others to do their work come from all around the world. In the 2015 MyMaster scandal in Australia, hundreds of students who were enrolled in more than a dozen universities paid a total of at least Aus\$160,000 (US\$108,000) to a 'service' that provided ghost-written essays and responses to online tests. In 2018, YouTube stars on more than 250 channels received money for promoting a cheating service called EduBirdie. Similar companies have been uncovered in the United States and elsewhere. Scientists should not deceive themselves: they are not immune.

Academics call this 'contract cheating'. My colleagues and I have assembled what is, to our knowledge, the largest data set on the topic – with responses from some 14,000 students and 1,000 teachers across 8 Australian universities. We found that roughly 6% of students have engaged in the practice; that most who cheat do so more than once; and that both post- and undergraduate students engage in it. Cheating is not new, but the proliferation of commercial, online services in the past 5–10 years has made it easier than ever.

And cheating is becoming increasingly normal. Since the 1990s, universities around the world have reimagined themselves as commercial enterprises that promote educational 'products' to student 'consumers'. In 2017, a commentator likened the brash marketing strategies of some UK universities to the advertising of shampoo, and hundreds of academic papers have openly criticized the 'marketization' of higher education. It's no wonder students opt to take the most convenient route to an academic credential – just as they would shop around for any other deal. In our survey, more than one-third of teachers specifically blamed contract cheating on the commercialization of higher education.

Data from student surveys uncovered three factors associated with contract cheating: speaking English as a second language; thinking that there are lots of opportunities to cheat; and dissatisfaction with the educational environment. Two trends – dwindling teaching resources and lower linguistic and academic standards for admission – contribute to the situation. And although little research has been done on the frequency of this phenomenon in scientific researchers, those driven to outsource their written work as undergraduates will probably be tempted to do so as academics under pressure to 'publish or perish'.

A cursory Google search for 'ghost-writing services for researchers' identifies thousands of services offering complete dissertations, grant applications, conference papers and journal articles. Outfits selling authorship on research

  
**Scientists should not deceive themselves: they are not immune."**

**Tracey Bretag** is an associate professor in the School of Management at the University of South Australia in Adelaide. e-mail: tracey.bretag@unisa.edu.au

publications have been uncovered in China and Iran, and the market for ghost-written PhD dissertations is reportedly booming in Ukraine and seems healthy in Australia.

We need to recognize that contract cheating is not just the responsibility of individual students, teachers or institutions. It is a systemic issue. Government funding agencies, regulatory authorities and leaders in higher education must tackle it.

Some have made a good start. In 2018, New Zealand successfully prosecuted commercial cheating service Assignments4U, which paid NZ\$2.1 million (US\$1.3 million) in an out-of-court settlement and closed down. In April 2019, the Australian Department of Education introduced a draft bill targeting commercial services that advertise or provide unauthorized assistance to students. It is expected to become law next year. Ireland has a similar law on its books. Such laws send a clear message. Cheating is not just unethical, it is illegal – and it has consequences. Laws hold accountable the stakeholders that are essential for contract cheating over which educational institutions have no control.

I feel that another important strategy is to reduce demand from within. Our team has found shockingly little concern in academics, including cheating students, non-cheating students and senior decision makers. They think that cheaters are just hurting themselves and are not damaging the community.

A radical shift in rhetoric would help individuals see the value of actually doing their work. Institutions need to stop treating education as a product and refrain from determining the value of research by the amount of funding received or the number of papers produced. Instead, they should focus on building academic cultures that are committed to integrity and that place abiding faith in the value of knowledge creation.

We cannot simply tell people not to cheat. We must provide support so that students feel capable of completing their assignments. That includes ensuring that institutions have appropriate language requirements for admission and allocating appropriate resources for teaching and learning.

Contract cheating is also a threat to public safety. It is not difficult to imagine how doctors, engineers and social workers who have outsourced their learning could pose a risk. This practice even threatens the common understanding of scientific facts – a big concern in the 'fake news' age. A large number of researchers purchasing their theses, publications and qualifications would endanger the credibility of science.

To defend itself, the scientific community must recognize that contract cheating is not an isolated problem caused by 'bad apples'. It is an attack on core academic values that necessitates stronger leadership from government departments, funders, regulators and educational institutions. This threat requires a collective response.

# News in brief

Trendwatch

## TRUMP REVIVES WHITE HOUSE SCIENCE COUNCIL

US President Donald Trump has resurrected the President's Council of Advisors on Science and Technology (PCAST), more than two years after he took office. Trump signed an executive order on 22 October appointing seven members to the panel. All but one come from industry, rather than academia. The lone university researcher on the panel so far is Birgitta Whaley, director of the Quantum Information and Computer Center at the University of California, Berkeley.

The four other members of the panel who have PhDs in scientific fields include H. Fisk Johnson, the chief executive of cleaning-products firm S. C. Johnson in Racine, Wisconsin; Dario Gil, director of IBM Research in Yorktown Heights, New York; Sharon Hrynkow, senior vice-president for medical affairs at Cyclo Therapeutics in Gainesville, Florida; and A. N. Sreeram, chief technology officer for the chemical firm Dow in Midland, Michigan.

The remaining members of the panel are Catherine Bessant, chief technology officer at Bank of America in Charlotte, North Carolina, and Shane Wall, chief technology officer at Hewlett Packard and global head of HP Labs in Vancouver, Washington.

PCAST will be led by Kelvin Droegemeier, director of the White House Office of Science and Technology Policy and Trump's science adviser.



## Fresh push for 'failed' Alzheimer's drug

A drug for treating Alzheimer's disease that the biotechnology company Biogen left for dead in March may get another chance to prove itself.

Biogen, in Cambridge, Massachusetts, announced on 22 October that it would seek approval from the US Food and Drug Administration (FDA) for its drug aducanumab to treat early-stage Alzheimer's disease. Biogen's share price, which plummeted in March following news that the drug had failed, shot up again after the latest announcement.

The company had halted development of aducanumab, an antibody that targets deposits of amyloid- $\beta$  protein in the brain, because an early analysis of clinical-trial results suggested that it had no significant effects on clinical symptoms such as memory loss and disorientation. But Biogen has since evaluated new data from the same studies. These showed that, for a subset of patients, high doses of aducanumab given for an extended period significantly slowed cognitive decline.

Scientists caution that FDA approval is not guaranteed.

SOURCE: NASDAQ



## WANTED: TOUGHER LAWS TO PROTECT AUSTRALIAN SPECIES

More than 240 conservation scientists have called on the Australian government to strengthen the country's environment-protection laws to halt habitat destruction and species extinctions.

In an open letter to Prime Minister Scott Morrison, the scientists wrote that Australia is in the middle of an "extinction crisis", citing evidence that in the past two centuries the continent has lost more than 10% of its 273 native land-mammal species. Two mammal and one reptile species have died out in the past decade. In addition, more than 1,800 plants and animal species are listed as threatened with extinction.

The scientists' call comes as a review of the national environment legislation starts this week. The Environment Protection and Biodiversity Conservation (EPBC) Act, established in 1999, must be reviewed at least once a decade.

"Our current laws are failing because they are too weak, have inadequate review and approval processes, and are not overseen by an effective compliance regime," the scientists wrote. Since 2000, more than 7 million hectares of habitat that could have housed threatened species has been wiped out.

Environment minister Sussan Ley said that the review of the EPBC Act will encourage submissions, such as the letter, as well as input from agricultural and industrial perspectives. "The government is investing significantly in environmental restoration and land care programs to promote biodiversity and safe havens for our native species," Ley said.



## CANADIAN KIDS SUE GOVERNMENT OVER CLIMATE CHANGE

A group of children and young adults filed a lawsuit on 25 October alleging that the Canadian government has violated their constitutional rights by promoting and enabling fossil-fuel development in spite of acknowledged risks from global warming.

Fifteen people aged between 10 and 19 filed the lawsuit in federal court, arguing that climate change will impinge on their right to "life, liberty and security". The lawsuit also argues that climate change will interfere with basic equality rights, given that the most severe effects of climate change will be borne by future generations.

"The courts must hold this generation to account for harms that are being done to the next," says Chris Tollefson, co-counsel for the plaintiffs and a specialist in environmental law at the University of Victoria in Canada.

Ira Reinhart-Smith, a 15-year-old plaintiff from Caledonia, Canada, got involved with climate activism last year. "My generation and generations to come are going to be exposed to things that the world has never been exposed to before," he says.

The lawsuit is the latest in a series of legal challenges aiming to force governments around the world to act on climate change.

## STARCRAFT II AI OUTPLAYS TOP-RATED GAMERS

Players of the video game StarCraft II faced an unusual opponent this summer. An AI called AlphaStar – built by Google's artificial intelligence firm DeepMind – was unleashed on the game's European servers, where it achieved a grandmaster rating, placing it within the top 0.15% of the region's 90,000 players. This result, reported on 29 October (O. Vinyals *et al.* *Nature* <https://doi.org/10.1038/s41586-019-1724-z>; 2019), shows for the first time that an AI can play at the highest levels of StarCraft II, which poses different challenges from other games AIs have aced.

StarCraft II players wage futuristic warfare by controlling human and alien armies in real time. Experienced players multitask by managing resources, executing complex combat manoeuvres and out-strategizing their opponent.

AlphaStar uses similar machine-learning techniques to AIs that DeepMind built to play chess and Go. They rely on neural networks, which learn to recognize patterns, and could be applied to other AI problems, such as self-driving cars.

AlphaStar crushed low-ranking opponents and amassed 61 wins out of 90 games against top-rated players, although it wasn't able to beat the best human player in the world.



## SNIPERS IN SOUTH KOREA TARGET BOARS WITH DEADLY VIRUS

South Korea has mobilized military snipers and drones along the demilitarized zone between itself and North Korea to stave off wild boars carrying African swine fever.

Cases of the highly contagious and lethal virus in pigs started to appear in South Korea last month. Authorities there have confirmed 15 cases in wild boars and 14 in domestic pigs on farms near the border with North Korea. There is no vaccine or treatment for the virus.

The nation has culled more than 150,000 pigs so far to contain the spread, says the agriculture ministry. The defence ministry says it has mobilized civilian teams of hunters, as well as military snipers, to take down wild boars near the North Korean border, according to South Korean media.

The Food and Agriculture Organization of the United Nations says that at least ten countries in Asia have ongoing outbreaks of the virus, which has wiped out millions of pigs.

The disease's spread comes as scientists in China reported on 17 October that they have obtained the most detailed picture yet of the virus's structure, which could assist vaccine development (N. Wang *et al.* *Science* <http://doi.org/ddbp>; 2019).

# News in focus



TIMOTHY ALLEN/GETTY

Archived weather data are helping to produce more-accurate climate predictions, such as the extent of rainfall in Mali.

## SCIENTISTS STRUGGLE TO ACCESS AFRICA'S HISTORICAL CLIMATE DATA

Better climate predictions require Africa's weather agencies to open up their archives. But commercial concerns and a lack of trust are holding them back.

By Linda Nordling

**F**or principal meteorologist Griefy John Stegling, the storerooms at Botswana's national weather-service headquarters in Gaborone hold a rare treasure: floor-to-ceiling shelves containing boxes of old notebooks with carefully recorded weather observations going back more than a century.

Such records offer clues not only to the country's past, but also to the future of its climate. Like most African countries, Botswana is ill served by global climate models, because predictions are based on patchy records of key variables such as temperature, humidity and atmospheric pressure (E. Archer *et al. Biodivers. Ecol.* **6**, 14–21; 2018).

"Historical climate data over Africa are very valuable for understanding climate variability and trends," says Chris Taylor, a meteorologist at the Centre for Ecology and Hydrology in Wallingford, UK, who studies African climate trends.

In 2017, Taylor and his team found that climate change will increase extreme rainfall in the Sahel, a semi-arid region south of the Sahara Desert (C. M. Taylor *et al. Nature* **544**, 475–478; 2017). A crucial part of their study involved cobbling together historical records – some of them "locked away in cupboards" – from different national weather services, Taylor says. "Having a historical baseline is a prerequisite for understanding how intense rainfall is changing," he says.

Since 2015, the World Meteorological Organization in Geneva, Switzerland, and Germany's weather service, Wetterdienst, have provided training and equipment to help Botswana digitize and share its historical climate data. But because there are no dedicated staff members, progress has been slow. Of 2 million records, only 100,000 have been processed. "If we had more manpower, it would go much faster," Stegling says.

Whereas Botswana is making some progress, in other meteorological offices across Africa, millions of records are mouldering in cardboard boxes or languishing on obsolete technology. Digitization efforts have been held up because of concerns that giving researchers free access to the data will prevent such offices



from making money by selling the information.

The South African Weather Service (SAWS) has turned down offers by the International Data Rescue (I-DARE) project to help digitize historical climate data because the agency wants to be able to sell its data. “If unrestricted access to the National Climatological Databank, of which SAWS is the custodian, is allowed, SAWS might not be able to deliver on its commercial mandate,” a spokesperson told *Nature*.

Similar concerns are holding up the digitization of 2 million surface observations – including temperature, rainfall and humidity – from 48 African countries. Those data are stored at the African Centre of Meteorological Applications for Development (ACMAD) in Niamey, Niger.

“The private sector is progressively being involved in climate-services delivery,” says ACMAD director-general Andre Kamga Foamouhoue, and this sometimes creates conflicts of interest with government agencies looking to commercialize data.

Many of the data-rescue requests come from initiatives led by individuals or institutions from Europe or the United States, says Jane Olwoch, executive director of the Southern African Science Service Centre for Climate and Land Management, a regional climate research centre in Windhoek, Namibia.

And that can be a problem because institutions in African countries aren’t sure how they will benefit if the data expertise comes from outside the continent. She hopes that data-rescue efforts fronted by her own organization, in Angola and Botswana, will be viewed with less suspicion because the organization is backed by four southern African governments and has local headquarters and staff, even though much of the funding comes from the German government.

### Recovering old records

Not all of Africa’s climate records are in Africa, however. Many of the oldest ones were collected by professional and amateur meteorologists who came to Africa from Europe during colonial times. Stefan Grab, a geographer at the University of the Witwatersrand in Johannesburg, South Africa, says that, paradoxically, these records can be easier to access than local ones.

South Africa has the Southern Hemisphere’s longest uninterrupted weather observations, recorded at the astronomical observatory in Cape Town. It was thought that these data stretched back to 1841, but Grab, who leads South Africa’s data-rescue efforts, knew that astronomers had been in the Cape since the 1830s. So he contacted staff at the Royal Greenwich Observatory in London, who directed him to the archives at the University of Cambridge, UK. “Lo and behold, they found the earliest records, which go back to 1834,” he says.

ACMAD’s Kamga Foamouhoue says that

weather agencies must be persuaded of the benefits of mining, and then sharing, historical data with other scientists, and that the biggest benefit is more-accurate climate predictions.

“Anything that’s really old, like from the nineteenth century, is extremely valuable,” Grab emphasizes. “It’s worth far more than gold and diamonds.”

## CANADIAN RESEARCHERS RELIEVED AS TRUDEAU EKES OUT ELECTION WIN

Government’s need to rely on progressive parties bodes well for climate policies and science funding.

By Brian Owens

**C**anada’s Prime Minister Justin Trudeau won a second term on 21 October, but his Liberal Party lost its majority in parliament. The Liberals and the Conservative Party were locked in a dead heat leading up to election night, and the outcome of the closely fought race brought a sense of relief for many researchers in Canada.

Apart from climate change, science issues didn’t feature prominently in the election campaign. But researchers are hopeful that several of the parties that won seats in parliament will work together to pass climate-change policies and increase science funding.

The left-leaning Liberal Party – which has had a mixed, but generally positive track record with respect to science – won 157 seats in parliament, 13 shy of a majority. The Conservatives won the second-largest number of seats at 121. This means that the Liberals will need to rely on other minority parties in parliament, specifically the New Democratic Party (NDP) and the Green Party, to help pass legislation. Such policies could include the Liberal proposal of

reaching net-zero carbon emissions by 2050.

Cathleen Crudden, a chemist at Queen’s University in Kingston, Canada, says that there was a real possibility that the Conservatives would win, and that they would undo the progress in science made over the past four years under the Liberal government. “What the Liberals started under their last mandate now has an opportunity to continue,” she says.

The government’s reliance on minority parties such as the NDP could be especially helpful to scientists when it comes to issues such as funding, says Molly Sung, a chemist at the University of Toronto in Canada who supported the NDP in the election. During the campaign, the NDP committed to boosting funding for basic research by Can\$80 million (US\$61 million) per year, she says. Sung thinks that their influence in a minority government could result in more money for science.

The NDP, along with the Liberals and the Greens, also supports the existing science-integrity policy that ensures government researchers can speak freely about their work. And all three parties have expressed interest in keeping the science-adviser position created by Trudeau in 2017.

Despite the agreement between the three progressive parties on science and climate issues, analysts also expect some friction. The NDP and the Greens say that they are open to cooperating with the Liberals. But they will probably demand concessions for their support – such as more-ambitious climate targets or increases in the country’s carbon tax – that the Liberals might not be comfortable with.

It’s important that researchers keep working to focus the government’s attention on science issues, says Katie Gibbs, executive director of the campaign group Evidence for Democracy in Ottawa. Minority governments tend to be messy as different interests jockey for power, so strong advocacy for science will be especially crucial, Gibbs says. “The science community is going to have to work hard and be loud to ensure these concerns don’t fall off the agenda.”



Justin Trudeau leads Canada’s Liberal Party.

COLE BURKSTON/GETTY





Historians want to build a virtual archive for Venice, pictured here in the eighteenth century.

# VENICE 'TIME MACHINE' PROJECT SUSPENDED AMID DATA ROW

Disagreements between international partners leave plans to digitize the Italian city's history in limbo.

By Davide Castelvecchi

**L**ike the city itself, an ambitious effort to digitize ten centuries' worth of documents that record the history of Venice is at risk of sinking. Two key partners have suspended the Venice Time Machine project after reaching an impasse over issues surrounding open data and methodology. The State Archive of Venice and the Swiss Federal Institute of Technology in Lausanne (EPFL) say they have had to pause data collection, and the archive's director has raised questions about the usability of the information that has already been collected.

The project sought to digitize documents that stretch over 80 kilometres of shelves in the state archive. These record the minutiae of the city's administration – from financial transactions to citizens' addresses and family connections – during its heyday in the Middle Ages and the Renaissance as a republic that dominated trade in the eastern Mediterranean for centuries. Many are written in Latin or the Venetian dialect, and have never been read by modern historians.

The goal was to make this information freely available online to researchers worldwide. The project also aimed to push the state-of-the-art in text-recognition technology for handwritten documents, using machine learning to automatically read millions of pages and tag their contents so that historians could perform quick searches.

The project was launched as a collaboration between EPFL, the State Archive of Venice and the Ca' Foscari University of Venice. In 2014, all three organizations signed a non-binding memorandum of understanding on how the work would be conducted.

However, the original agreement left out crucial details on the research protocols, according to a 19 September press release from the archive announcing the suspension. In particular, it didn't specify the type of licensing that would regulate researchers' use of the digitized data – which must also comply with Italian law, says the archive's current director, Gianni Penzo Doria. He adds that he tried to jump-start negotiations for a detailed contract after he took up his post in August, but the two sides came to an impasse. The decision to halt

the project was inevitable, he says, and mutual.

But on 23 September, EPFL issued its own sharply worded press release saying that the archive had suspended the project unilaterally, and that EPFL was surprised to learn of the decision from the archive's website. "I think it's essentially a misunderstanding," says Frédéric Kaplan, a computer scientist at EPFL and director of the Venice Time Machine. He adds that the disagreement could potentially have been resolved by face-to-face meetings between the collaborators, but that so far all discussion had been by teleconference.

## 'Useless' files

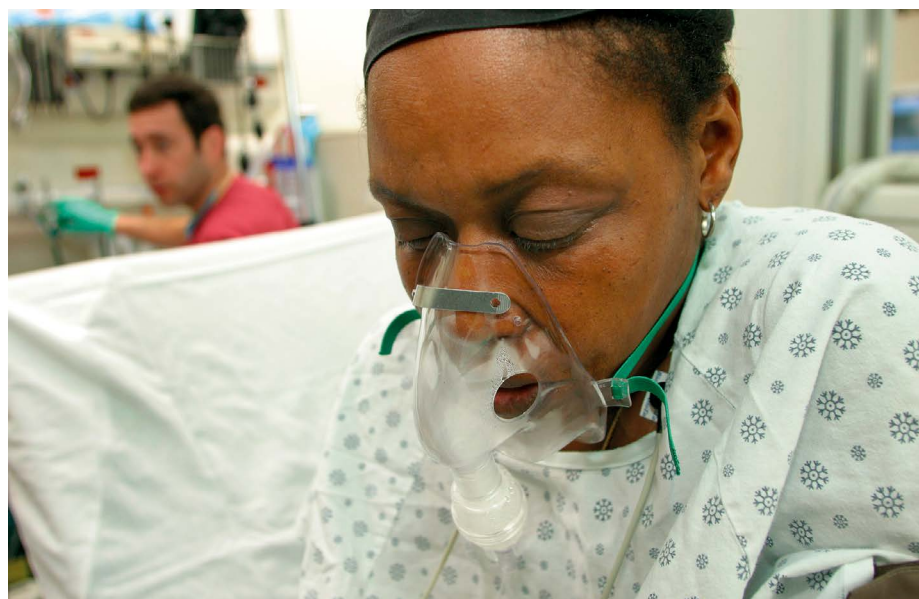
Meanwhile, the fate of the 8 terabytes of data accumulated over the past 5 years – from around 190,000 documents – is unclear. Penzo Doria claims that from the point of view of archival science, "these files are useless", because their digitization did not follow guidelines set by the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) project.

These guidelines mandate the scrupulous recording of information that certifies the provenance of each document, and require that a record of such information be kept in the metadata that comes with each file. This serves as a sort of electronic signature that ensures the long-term preservation and validation of a digital file. According to Penzo Doria, the EPFL researchers who made the scans did not document how they collected such information – or, if they did, they didn't share such documentation with collaborators at the archive.

Kaplan says the researchers did collect metadata, but based their methodology on a the International Standard Archival Description guidelines from the International Council on Archives. He says that EPFL researchers followed procedures established by the state archive's own staff. Kaplan also says that he provided documentation on the metadata in an e-mail to Penzo Doria's predecessor, Giovanna Giubbini, in February 2019. Penzo Doria and Giubbini both told *Nature* that they never received this documentation.

Raffaele Santoro, who was director of the State Archive of Venice in 2014 when the memorandum on the Time Machine project was signed, says that he doesn't know precisely how workers collected the metadata, but he assumes that they are scientifically valid because the archive's own staff were closely involved in the process. To make the documents that have already been digitized compliant with additional standards, one could simply add more information to the metadata, he says, "without any need to do it all over again".

Kaplan says he is hopeful the project can get back on track if the two sides meet to discuss new terms in person. "EPFL sincerely hopes the meeting will happen soon," he says.



Black people were less likely than white people to be sent for personalized care, a study found.

# MILLIONS AFFECTED BY RACIAL BIAS IN HEALTH-CARE ALGORITHM

Study reveals widespread racism in decision-making software used by US hospitals.

By Heidi Ledford

**A**n algorithm widely used in US hospitals to allocate health care to patients has been systematically discriminating against black people, a sweeping analysis has found.

The study, published in *Science* on 24 October, concluded that the algorithm was less likely to refer black people than white people who were equally sick to programmes that aim to improve care for patients with complex medical needs (Z. Obermeyer *et al. Science* 366, 447–453; 2019). Hospitals and insurers use the algorithm and others like it to help to manage care for about 200 million people in the United States each year.

This type of study is rare, because researchers often cannot gain access to proprietary algorithms and the reams of sensitive health data needed to fully test them, says Milena Gianfrancesco, an epidemiologist at the University of California, San Francisco, who has studied sources of bias in electronic medical records. But smaller studies and anecdotal reports have documented unfair and biased decision-making by algorithms used in everything from criminal justice to education and health care.

“It is alarming,” says Gianfrancesco of

the latest study. “At the same time, it’s not surprising.”

Ziad Obermeyer, who studies machine learning and health-care management at the University of California, Berkeley, and his team stumbled across the problem while examining the impact of programmes that provide additional resources and closer medical supervision for people with multiple, sometimes overlapping, health problems.

When Obermeyer and his colleagues ran routine statistical checks on data they received from a large hospital, they were surprised to find that people who self-identified as black were generally assigned lower risk scores than equally sick white people. As a result, the black people were less likely to be referred to the programmes that provide more-personalized care.

The researchers found that the algorithm assigned risk scores to patients on the basis of total health-care costs accrued in one year. They say that this assumption might have seemed reasonable because higher health-care costs are generally associated with greater health needs. The average black person in the data set that the scientists used had similar health-care costs to the average white person.

But a closer look at the data revealed that the average black person was also substantially sicker than the average white person, with

a greater prevalence of conditions such as diabetes, anaemia, kidney failure and high blood pressure. Taken together, the data showed that the care provided to black people cost an average of US\$1,800 less per year than the care given to white people with the same number of chronic health problems.

The scientists speculate that this reduced access to care is due to the effects of systemic racism, ranging from distrust of the health-care system to direct racial discrimination by health-care providers.

And because the algorithm assigned people to high-risk categories on the basis of costs, those biases were passed on in its results: black people had to be sicker than white people before being referred for additional help. Only 17.7% of patients that the algorithm assigned to receive extra care were black. The researchers calculate that the proportion would have been 46.5% if the algorithm was unbiased.

When Obermeyer and his team reported their findings to the algorithm’s developers – Optum of Eden Prairie, Minnesota – the company repeated their analysis and got the same results. Obermeyer is working with the firm without salary to improve the algorithm.

He and his team collaborated with the company to find variables other than health-care costs that could be used to calculate a person’s medical needs, and repeated their analysis after tweaking the algorithm accordingly. They found that making these changes reduced bias by 84%.

“We appreciate the researchers’ work,” Optum said in a statement. But the company added that it considered the study’s conclusion to be “misleading”. “The cost model is just one of many data elements intended to be used to select patients for clinical engagement programs.”

Obermeyer says that using cost prediction to make decisions about patient engagement is a pervasive issue. “This is not a problem with one algorithm, or one company – it’s a problem with how our entire system approaches this problem,” he says.

## Examining assumptions

Correcting bias in algorithms is not straightforward, Obermeyer adds. “Those solutions are easy in a software-engineering sense: you just rerun the algorithm with another variable,” he says. “But the hard part is: what is that other variable? How do you work around the bias and injustice that is inherent in that society?”

This is in part because of a lack of diversity among algorithm designers, and a lack of training about the social and historical context of their work, says Ruha Benjamin, author of *Race After Technology* (2019) and a sociologist at Princeton University in New Jersey.

“We can’t rely on the people who currently design these systems to fully anticipate or mitigate all the harms associated with

ED KASHI/VII/REDUX/EYEVINE



automation,” she says.

Developers should routinely run tests such as those performed by Obermeyer’s group before they deploy an algorithm that affects human lives, says Rayid Ghani, a computer scientist at Carnegie Mellon University in Pittsburgh, Pennsylvania. That kind of auditing is more common now, he says, since reports of biased algorithms have increased.

He thinks that the results of these audits should always be compared to human

decision-making. Unpublished analyses by Ghani’s team have compared algorithms used in public health, criminal justice and education to human decision-making, and found that the machine-learning systems were biased – but less so than the people.

“We are still using these algorithms called humans that are really biased,” says Ghani. “We’ve tested them and known that they’re horrible, but we still use them to make really important decisions every day.”

gradient of same-sex attraction. The app cited the *Science* study and warned users that it did not predict same-sex attraction.

The researchers behind the *Science* study say that Bellenson’s app misrepresents their work. The test “is not grounded in science. It is not predictive. It won’t tell you anything”, says Benjamin Neale, a geneticist at the Broad Institute and an author of the *Science* analysis. He and his colleagues examined the DNA of around 475,000 people and found 5 genetic variations loosely correlated with people who said they’d had sex with someone of the same sex at least once. But none of the variants was so prevalent that the researchers could use them to predict a person’s sexual identity.

Neale sent a letter to GenePlaza on 14 October asking that it take down the app – or remove references to his study. The next week, Bellenson renamed the app ‘122 Shades of Gray’ and added a note explaining that the authors of the *Science* study weren’t affiliated with the project. He says that because the app has always warned users that it is not predictive, it does not misrepresent the study.

But the chorus of angry scientists on Twitter grew louder. Some echoed Vitti’s concern that the app could be abused. In his petition, Vitti noted that Bellenson lives in Uganda, where gay sex is punishable by life in prison. Vitti worried that, regardless of the app’s scientific flaws, Ugandan authorities could get hold of a person’s results and use them as evidence of sexual preferences.

Bellenson says that there are much simpler ways of discovering a person’s sexual preference, such as looking at their social-media accounts. “The idea that a government would need a DNA test to figure out if someone is gay is ridiculous,” he says.

# ‘GAY GENE’ APP PROVOKES FEARS OF A GENETIC WILD WEST

## Debate highlights broader concerns about apps that use the results of direct-to-consumer genetic testing.

By Amy Maxmen

Joseph Vitti’s stomach turned when he opened a link an acquaintance had sent him. It took him to an app called ‘How Gay Are You?’ that purported to gauge a person’s level of attraction to others of the same sex, according to their genes.

The app’s creator, Joel Bellenson, a US entrepreneur living in Kampala, Uganda, based the test on the findings of a massive study on the genetics of same-sex sexual behaviour – even though the analysis, published in *Science* in August, concluded that a person’s genes cannot predict their sexuality (A. Ganna *et al. Science* 365, eaat7693; 2019).

Vitti, a computational geneticist at the Broad Institute in Cambridge, Massachusetts, thinks the app was misleading – even dangerous. “There are vulnerable queer people all over the world,” says Vitti, “and this app stands to hurt them.” On 11 October, he started an online petition to remove the test. Within two weeks, more than 1,660 people had signed it.

Bellenson says that the idea his test could endanger people is an “absurd scenario” and notes that the test also included a warning that it could not predict same-sex attraction.

But the furor over his app highlights a growing problem in the field of genetics. Researchers conduct statistically sophisticated analyses of hundreds of thousands of genomes, searching for associations between genetic variations and diseases, behaviours or other characteristics. Anyone can take the variations identified by such studies, strip them of caveats and nuance, and market a simple genetic-interpretation tool online.

Scientists and genetic counsellors say that

these unregulated tools can harm individuals and society, causing anxiety, unnecessary medical expenses, stigmatization and worse. “It’s the Wild West of genetics,” says Erin Demo, a genetic counsellor at Sibley Heart Center Cardiology in Atlanta, Georgia. “This is just going to get harder and harder.”

Bellenson posted his app on GenePlaza, an online marketplace for DNA-interpretation tools, in early October. For US\$5.50, a person could upload their genetic data – as supplied by consumer DNA sequencing companies such as 23andMe of Mountain View, California – and the app would place them along a



Millions of people have had their DNA sequenced by consumer genetic-testing companies.

## News in focus

On 24 October, GenePlaza co-founder Alain Coletta removed the app from his platform. He and Bellenson both say they did not intend to hurt anyone by making the app available. And they echo other creators of third-party tools that interpret DNA sequencing data, who say that even if their tests aren't predictive, they encourage public engagement in science. "It may not be much better than a horoscope or a tarot-card reading, but at least it lets bioinformatics be something fun," Bellenson says.

This argument concerns genetic counselors, who have seen a surge in the number of people seeking help for conditions that third-party tools have identified in their DNA – often inaccurately.

Tens of millions of people worldwide have now had their DNA sequenced by direct-to-consumer companies. But these sequencing companies only highlight certain genetic associations. If customers want more information, they can download their raw genetic data from these firms' sites for further exploration.

Up to 62% of customers ultimately upload their genetic data to third-party websites for a small fee, a study published in August found (T. Moscarello *et al. Genet. Med.* **21**, 539–541; 2019). GenePlaza, for example, offers DNA-interpretation apps that purport to assess intelligence, neuroticism and taste perception. Other websites advertise services that use

DNA to explore a person's ancestry, disease risk, ideal romantic partner, fondness for marijuana, nutritional needs, sleep habits and more.

In 2015, 23andMe learnt that its customers could feed their DNA data directly from 23andMe's servers into a secondary application associated with white supremacists

**"Scientists have a responsibility to describe the human condition in a more nuanced and deeper way."**

that evaluated a person's degree of European ancestry. 23andMe shut down the app's access to data on its servers. The company went further last year by restricting direct access to its data to select collaborators.

23andMe also warns customers who download their data it cannot ensure the accuracy of third-party interpretation tools. Developers of these tests might base them on genome-wide analyses that find weak correlations, or associations that have been contradicted by additional analyses.

But Vitti thinks scientists should bear more responsibility for how their results are used – especially now that geneticists are delving deeper into social and behavioural traits,

such as links between a person's DNA and their political persuasion.

He argues that ethical review boards should assess whether the benefits of such studies outweigh the potential for harm. Genome-wide analyses are not scrutinized to the same degree as research on individuals because the data they rely on are pooled and anonymized. But the How Gay Are You? app illustrates how such analyses could lead to harmful outcomes, Vitti says.

Despite his distaste for the app based on his study, Neale says research must go on. "Scientists have a responsibility to describe the human condition in a more nuanced and deeper way," he says.

But Sarah Nelson, a geneticist at the University of Washington, Seattle, who has studied third-party interpretation tools, worries that her peers aren't fully aware how difficult their studies can be for the public to understand. Even if researchers take pains to explain that their genome-wide analyses aren't predictive, she says, companies can still use the science as they please – and the barrier to entry is low.

Indeed, Bellenson says he whipped together his app in a weekend. He knew enough about genetics and computer programming to write an algorithm, and find a home for it online. "Genetics and bioinformatics is so mature," he says. "Academia can no longer control it."

# Promoting diversity

Read our series on diversity within the sector  
[go.nature.com/diversityinscience](https://go.nature.com/diversityinscience)

nature careers

A80550





A tiger and her cubs leave India's Bandhavgarh National Park in search of prey, which is scarce inside the reserve.

# TIGER TROUBLE





**India is trying to save its tiger population, but researchers question the country's long-term plans to protect this endangered species.**  
**By Gayathri Vaidyanathan**

**C**entral India – The Maruti Gypsy 4×4 sped along a jungle track, jolting us out of our seats. We had signed up for a wolf safari, but the trip leader had another quarry in mind. The vehicle barrelled towards a pungent smell on a hillside – a fresh tiger kill.

The forest guide spoke to one of his colleagues in a different vehicle and then barked at our driver to rush towards a nearby meadow. A tigress and four cubs are at a watering hole just beyond our sights, he said.

A full Moon rose, and revealed an ink-blue landscape. Handheld lights were banned, so visibility was at 3 metres. The phone rang, and the guide instructed the driver, who raced on a rollercoaster route back to the kill site. No tiger. We dashed back to the meadow, a second vehicle in hot pursuit. It felt ugly, like a hunt.

Two circuits later, the Moon was high over the meadow when we were beckoned once more back to the kill site. We raced there to find four Gypsies, the drivers using their headlights to sweep the hillside. Another vehicle banged into ours. Our guide cursed. Then silence, as the drivers shut off the engines. Tourists stood on seats, peering through telephoto lenses.

Footsteps rustled dead leaves, and the drivers switched on the high beams. There sat two tigers, larger than life as wild tigers are. These were no cubs; they were male adolescents. Camera shutters clicked. Minutes later, the animals got up and disappeared into the darkness.

Two hundred years ago, tens of thousands of tigers (*Panthera tigris*) roamed India and 29 other nations, from the Indonesian swamps to the Russian taiga. There were once Balinese, Caspian and Javanese subspecies, all now considered extinct. Today, only six subspecies remain. The International Union for Conservation of Nature (IUCN) estimated in 2014 that there are only about 2,200 to 3,200 individuals in the wild, placing the animal on the organization's endangered list. About 93% of the tiger's historic range has emptied owing to habitat loss, poaching and depletion of prey.

The spectre of a world without tigers led 13 nations to meet in 2010 in St Petersburg, Russia, where they declared that they would double their wild tiger numbers by 2022. But all except India, Nepal and Bhutan are struggling to save their tigers, even in protected reserves.

Against this backdrop, India is the beacon. It has roughly two-thirds of the world's tigers in less than one-quarter of their global range. In 2019, it has invested 3.5 billion rupees (US\$49.4 million) in tiger conservation, including relocating villages outside protected areas. And it has built the world's largest animal underpass to funnel tigers safely beneath a highway.

About 3% of the spending on tigers is flowing to government-sponsored science. Government scientists are studying all aspects of the animal, and are heading a large tracking study to understand tiger behaviour.

STEVE WINDER/NATIONAL GEOGRAPHIC



The efforts have paid off, according to the government. It announced in July that the number of wild tigers in the country had doubled from 1,411 in 2006 to 2,967 today – meaning that India has met the St Petersburg target. Indian Prime Minister Narendra Modi declared that tiger conservation could go hand in hand with building roads, railways and homes.

But parse the country's tiger data, and the story becomes murky. The animals are increasingly becoming isolated in small reserves that prioritize tourism. If the cats leave the parks, the risks are rising that they will encounter humans and infrastructure, with tragic results for both the animals and people. Some scientists question whether tiger numbers in India have truly increased and are attempting to get a more accurate count of populations in specific areas. Other researchers are studying how to get people and the carnivores to coexist.

Saving tigers is difficult enough, but research efforts in India are made more challenging by an apparent antagonism between the actors involved. Some experts charge that government scientists sometimes present questionable evidence in support of state policies and hamper efforts by independent investigators. Such conflicts are routine in tiger conservation globally, says John Goodrich, who heads the tiger programme at Panthera, a conservation organization in New York City.

"It's something that I've been incredibly frustrated with," he says. "We all have all this data, all this knowledge that we need to be sharing."

## The national animal

Two hundred years ago, an estimated 58,000 tigers roamed India's lush, unbroken forests<sup>1</sup>. But centuries of hunting and

habitat destruction left fewer than 2,000 wild individuals by the 1970s. In 1973, the government declared the tiger India's national animal, banned hunting and set up a conservation scheme called Project Tiger. There are 50 reserves today under the programme, and about half are well managed, according to a government assessment. But the reserves are small, averaging less than 1,500 square kilometres – much smaller than many protected areas in Africa.

These are unfavourable conditions for the solitary tiger. Male Bengal tigers need a home range of about 60–150 km<sup>2</sup>, whereas females use about 20–60 km<sup>2</sup>. And tigers do not share easily, even with siblings or kids. So when a cub hits adolescence at about one and a half years, it begins roaming to find territory in which to live and hunt. If the tiger reserve is already full, it has two options: either push out an old or weak tiger and take over the space, or keep moving well outside the reserve until it finds unoccupied territory. It is thought that 70–85% of India's tigers are inside reserves.

These numbers are from India's tiger census. Every four years, an army of forest guards, conservationists and volunteers fans out over an area roughly the size of Japan and carries out a comprehensive census. It's a difficult task because tigers are elusive. The workers place camera traps in some parts of tiger reserves for about 35 days. Then they walk on foot, collecting sightings of tiger tracks, scat and signs of prey and human disturbance. This is called a sign survey. They send the data to scientists at the government-run Wildlife Institute of India (WII) in Dehradun, who identify individual tigers in photos from their unique stripe patterns and then estimate local tiger densities in reserves. They create a calibration model that

links the tiger densities to the collected signs, then input the sign-survey data into this model to derive nationwide numbers.

"Unless you know what you have and where you have it, you can't manage it," says Yadavendra Dev Jhala, who heads the tiger team at the WII and is responsible for the survey.

The latest census suggests that tigers are rebounding, and Modi celebrated a 33% increase in numbers since 2014.

But many scientists are sceptical. Ullas Karanth, director of the Centre for Wildlife Studies in Bengaluru, questions the sign surveys, which he says are collected by ill-trained workers who don't know how to do accurate counts – an accusation he based on his own experiences with the field workers. "The field protocols are deeply flawed," Karanth says. When I walked with forest guards doing surveys in a reserve in May, they said they felt pressured by local officials to record positive tiger signs and ignore signs of human disturbance.

Critics also argue that Jhala's team varies the census coverage every time. In 2018, they added 90 survey sites and 17,000 extra cameras. These types of differences make it difficult to compare census years and to say how India's tigers are faring, says Abhishek Harihar, a population ecologist with Panthera in Bengaluru.

Another point of contention is the data analysis, particularly the calibration model used to arrive at pan-India numbers. The description of the methodology and models used is "vague", and the resulting numbers have "higher uncertainties than are currently reported", says Arjun Gopalaswamy, a statistical ecologist and science adviser at the Wildlife Conservation Society in New York City. He has authored two studies critiquing the census method<sup>2,3</sup>.

Jhala refutes the criticisms about the accuracy of the census. He says there are safeguards to protect against bad data. Although the coverage has increased, he says the census is based on estimates of tiger density, so increasing the extent of the survey does not affect the trend calculations. He has published a study refuting the accusations<sup>4</sup>.

The best way to resolve the disagreement, argue scientists, would be if the WII released raw data and model information to ecologists for independent analysis. But Jhala says that releasing the geo-tagged data, even to scientists, could make the animals vulnerable to poaching – a claim that others dispute.

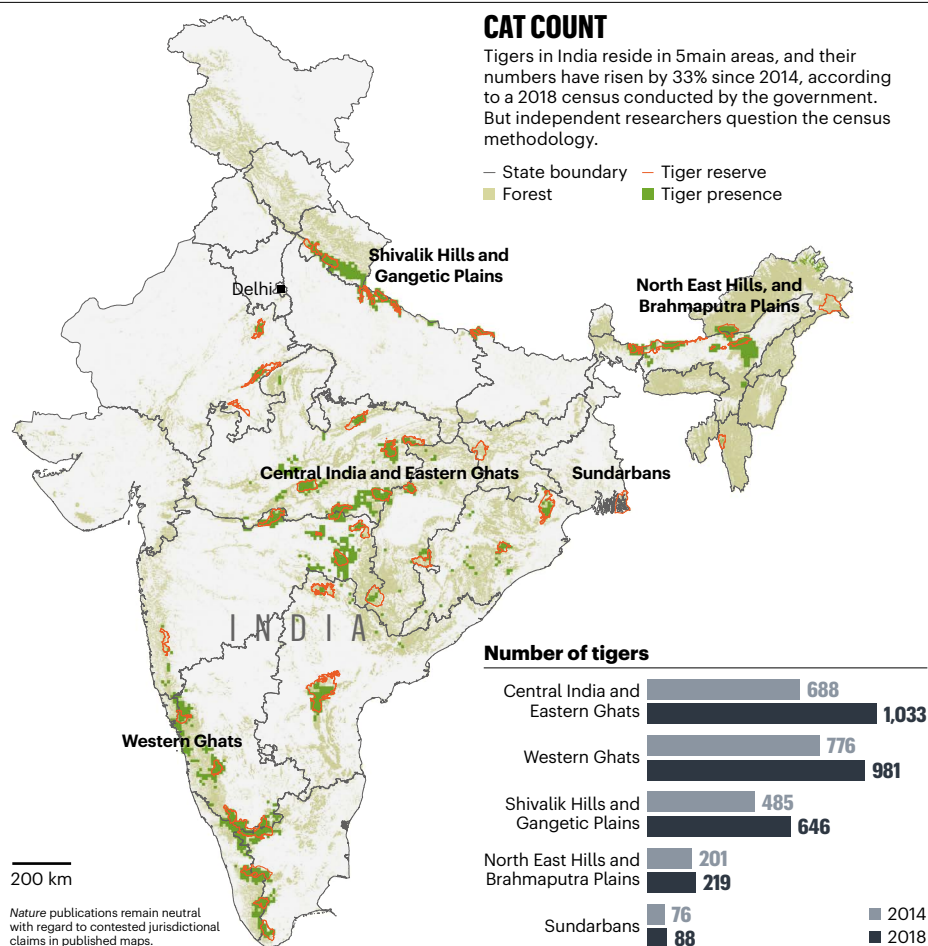
The result is that there is little consensus on India's tiger population and, more importantly, whether it is rebounding or has remained steady for many years. For now, scientists can say only that the animals are thriving in some places, but doing poorly elsewhere.

The biggest known conservation success is in central India, an area with 19 tiger reserves across 8 states. I travelled there in May with researchers from the Mumbai-based Wildlife Conservation Trust to see how India's best-kept



A tiger was killed trying to cross a train track near Jaldapara National Park, India.





tigers are faring.

The central Indian forests in Maharashtra state were brown and crackly under the 45 °C heat. Most trees had dropped their leaves for the dry season, reservoirs had dipped low and everyone was waiting for the monsoons.

The government says there are 1,033 tigers in central India, up 50% since 2014 (see 'Cat count'). That's more than one-third of India's tigers. The region attracts a proportionally high number of India's tiger scientists.

They have found that historically, tigers here have moved unhindered through forest corridors in search of territory, carrying precious new genes into distant populations. The central Indian tigers have high genetic variation, which should help them to adapt to environmental crises such as drought or disease<sup>5</sup>.

But the forest corridors in central India are fragmenting rapidly. Without roaming tigers, none of India's small reserve populations would be demographically viable in the long run, says Aditya Joshi, head of conservation research at the Wildlife Conservation Trust. Uma Ramakrishnan, an ecologist at the National Centre for Biological Sciences in Bengaluru, says that if infrastructure development in rural areas continues unabated, the genetic diversity of small populations could fall within a century.

The government might then have to shuttle tigers between reserves to maintain the gene

flow necessary for a population to stay healthy. "That will be pretty much like a zoo," she says.

In the worst-case scenario, tigers might get marooned in reserves and relatives might start breeding. These aren't vague fears. In the Ranthambore tiger reserve, a popular tourist attraction in northwest India, some 62 individuals, half of them descended from one matriarch, live in genetic isolation in a 1,115 km<sup>2</sup> area. Villages surround the reserve, and there are no other tiger populations nearby to seed new genes. Ramakrishnan and her colleagues have seen markers of inbreeding in the genomes of Ranthambore tigers<sup>6</sup>. In an unpublished study, they have detected regions of over a million base pairs of DNA without variation. In an average tiger, there are 500 variations in every million or so base pairs. If these stretches harbour deleterious alleles, the offspring could have reduced fitness, increasing the risk of local extinction, she says.

### Deadly highways

The day before the frenzied night-time chase in Pench tiger reserve, Milind Pariwakam, a road ecologist with the Wildlife Conservation Trust, and I drove there on a four-lane motorway called National Highway 44, or NH44 (also known as NH7). In a nation full of potholes, I appreciated the smooth road connecting two major cities and reducing travel time. But

Pariwakam says the road comes at a high cost.

A 65-kilometre section of the NH44 cuts through the tiger park, separating the core reserve from a forest corridor. Some 40 mammalian species, including tigers, use this landscape. So do 6,151 trucks, cars and motorcycles that race down the NH44 every day. And this is not the only road through Pench; there are 24 smaller roads and another highway – the NH6.

Roads kill millions of animals globally every year. And over time, busy roads become barriers to movement as some species learn to avoid them. Tigers, which prefer to stroll on paths rather than skulk through undergrowth, are attracted to roads and exhibit little fear of traffic. In the Russian Far East, home of the Siberian tiger, scientists looked at the impact of roads on 15 resident individuals. The roads carried 250 vehicles a day, a fraction of the traffic through Pench. The researchers found that tigers living in the area died sooner and had fewer offspring than did animals living in road-free areas<sup>7</sup>.

In 2008, Pariwakam and a group of non-governmental organizations sued the government to stop the expansion of the NH44 to four lanes. The fight lasted eight acrimonious years before WII scientists and the conservationists came to a compromise: underpasses that animals could use to walk safely beneath.

"What we always say is that conservation has to be affordable, it has to be sustainable," says Bilal Habib, a carnivore biologist who heads the central India tiger programme at the WII. "We are a developing nation."

Finished in 2018, the NH44 has 9 specially built underpasses, ranging in length from 50 to 750 metres, designed to allow animals to pass beneath the roads. These are the longest animal underpasses in the world, and the first in India. If evidence suggests they are effective, the government might deploy them in some of the 20,000 km of roads through wild spaces, Habib says.

But although the underpasses are excellent on paper, Pariwakam questions their efficacy. Since 2018, two leopards and one tiger have walked across the road rather than using an underpass and were hit and injured. As we inspected one structure, a 4×4 careened into view from a village access road and drove through the underpass to a service ramp leading to the highway. Pariwakam whipped out his phone and filmed the intrusion. "The villagers are using the shortcut to save a quarter kilometre," he said, seething. He has been urging the forest department to close off all access roads.

### Mistaken identity

This year, news about tiger deaths and fatal attacks on humans has popped up almost every week. As reserves have filled up, tigers are moving into the forest corridors that connect them – which are also used by people.

Tigress T49 was born in the corridor enclosing the Chandrapur district, outside the Tadoba Andhari tiger reserve, not too far from Pench. There are 155 people per square kilometre here, living in 600 villages that are slowly encroaching into forests. There are also 41 tigers, which is more than in half of India's protected reserves.

In December 2016, T49 had four cubs, named E1 to E4, in a culvert under a bridge. Villagers thronged on tractors and motorbikes to see the newborns.

Habib of the WII and his graduate student Zahidul Hussain were also interested in the cubs. Since 2013, Habib's team has radio-collared adolescents to understand the behaviour of tigers inside and outside reserves and to learn about the drivers of human–tiger conflict. They have collared 23 individuals so far, a small sample size. But this is still the largest telemetry, or tracking, study of tigers in the world. Their preliminary data are troubling. They suggest that non-reserve tigers move longer distances daily, perhaps to avoid humans and get around infrastructure. Consequently, they need 22% more food in an area already depleted by humans of wild prey. Habib says that of five tigers that left a reserve the team was monitoring, four died from walking into electrified wires.

In March 2019, the scientists collared E1, E3 and E4; E2 was shy and escaped, a trait that might serve her well amid humans. E1 was special. "As soon as you take your vehicle towards them, E1 is the first one to come to you," Hussain says. "She comes, sits there, curious about what is happening."

The adolescents were looking for territory, but roads, villages and the summer's sparse tree cover restricted their movement. E1 favoured a forest fringing a village.

On 6 April, an older woman went into the forest to collect flowers of the mahua tree, used to make liquor. As she crouched down, her posture made her look like small prey, researchers suspect. A tiger emerged without a sound and pounced. It dragged the woman 3 metres, then dropped her and disappeared.

There were two more human kills in three weeks. Hussain's data showed that E1 had been at all three kill sites, but none of the people was eaten, suggesting they were victims of mistaken identity; tigers generally don't eat humans. Scientists and the forest department are racing to understand ways to minimize such human–animal encounters. Some are using camera traps to warn villagers when tigers are in their vicinity. Others are exploring ways to train locals in alternative livelihoods so that they don't need to enter forests. Their efforts are urgent because the death toll is rising. Across central India, villagers have killed 21 tigers through electrocution, traps or poisoning since 2015. In Chandrapur alone, tigers have killed 24 people in the past 4 years.

In June, the forest department captured the tiger E1 and moved her to a wildlife rehabilitation centre, making her the ninth individual to be relocated since 2015. But it might be a temporary reprieve, as another tiger will probably take E1's territory.

## Scientific battles

Much like the animals they study, tiger scientists are fiercely territorial. Everyone except Karanth at the Centre for Wildlife Studies requested anonymity while speaking about politics because it could hinder their ability to do research.

Several scientists say there is a conflict of interest because government managers fund and oversee science as well as set policies regarding reserves. Karanth says managers

**"IT'S A NIGHTMARE WORKING IN WILDLIFE IN THIS COUNTRY."**

grant research permits more easily to scientists from the government-run WII than to independent scientists, unless the latter join government-led studies as junior partners. Independent observers also charge that government scientists sometimes rubber stamp government actions, whether or not they are scientifically sound.

"[The WII] seems to have completely bought in, they seem very biased," one scientist says. An example is the NH44 road project: although the WII initially recommended much larger overpasses to the government, it reworked its assessment to reduce costs and make it more palatable under pressure from government officials, according to a government report.

Most independent field initiatives have shut down, says Karanth. His 30-year-long study of tigers in southern India ended in 2017 because the local forest officials had been repeatedly interrupting or delaying his work – for example, by not allowing his assistants access to field sites. The union and state government officials ignored his complaints. "To get the permits was becoming very impossible," Karanth says.

"Sadly, I've realized I don't think I can impact policy," another scientist notes.

Government officials and researchers challenge those criticisms. Anup Kumar Nayak, member secretary of the National Tiger Conservation Authority (NTCA), India's tiger conservation and research coordinating body, says that his agency has permitted several research projects by non-government scientists and non-profit organizations. "Most

of the research projects are given to [the WII] because they are the technical wing of NTCA, and they've been doing research work on wildlife for a long time," he says. "In southeast Asian countries, they're a very reputed organization."

Nitin Kakodkar, who is the chief wildlife warden of Maharashtra and signs off on research permits in his state, disagrees that WII scientists are favoured or that the managers influence research. WII scientists, he says, are more knowledgeable about the permit requirement procedures than are independent scientists. And he contends that there is no favouritism in Maharashtra. "There are people who've been doing research in Maharashtra who are not from the Wildlife Institute of India."

Jhala of the WII says his team finds it easier to get permits because they work for the government, but not by much. The bureaucracy is difficult even for WII scientists, he says. "It's a nightmare working in wildlife in this country."

The government maintains a tight grip because the tiger is a symbol of national pride, researchers say. That exalted status – and rising revenues from the tourism industry around tiger safaris and luxury resorts – might be what eventually saves the tiger from extinction.

The Indian government has plans to expand tiger conservation. For example, India is going to increase the number of tiger reserves in coming years, says Nayak.

Although numbers are stagnating in other countries, the 'wild' tiger will probably survive in India, at least inside reserves, researchers say. The animal's fate outside reserves is more questionable. Older villagers don't mind the large carnivores in their midst, but the younger generation is more circumspect.

In Kurwahi, a village near Pench, a 'fat' tiger snatched away a calf tied outside an elderly woman's house in March. Hers was one of 17 village cattle killed by the animal. I asked her whether she was angry. She put her palms together, laughed and shook her head. "How can I be angry with a tiger?" she said.

Her son glanced hesitantly at the forest-department guard, who was standing nearby. Then he gathered up courage and said what other villagers had been demanding – that there needs to be a more permanent solution. "The authorities should remove the tiger."

**Gayathri Vaidyanathan** is a science journalist in Bengaluru.

1. Mondol, S., Karanth, K. U. & Ramakrishnan, U. *PLoS Genet.* **5**, e1000585 (2009).
2. Gopalaswamy, A. M., Delampady, D., Karanth, K. U., Kumar, N. S. & Macdonald, D. W. *Methods Ecol. Evol.* **6**, 1055–1066 (2015).
3. Gopalaswamy, A. M., Karanth, K. U., Delampady, M. & Stenseth, N. Chr. Preprint at bioRxiv <https://doi.org/10.1101/708628> (2019).
4. Qureshi, Q., Gopal, R. & Jhala, Y. *PeerJ* **7**, e7482 (2019).
5. Thatte, P., Joshi, A., Vaidyanathan, S., Landguth, E. & Ramakrishnan, U. *Biol. Conserv.* **218**, 181–191 (2018).
6. Armstrong, E. et al. Preprint at bioRxiv <https://doi.org/10.1101/696146> (2019).
7. Kerley, L. L. et al. *Conserv. Biol.* **16**, 97–108 (2002).



# A COMPLEX INHERITANCE

Researchers are finding links between people's genes and complex attributes such as how they vote or the time spent in school. The worry is that their results will be misconstrued. **By David Adam**



**T**he deep coal mine at the Yorkshire village of Kellingley closed in 2015 – the last of more than 1,000 such pits that once drove British industry. As the mines closed, the jobs went with them. Faced with economic and social decline, many people who could moved away.

Geneticist Abdel Abdellaoui has never been to Kellingley or any of the United Kingdom's other former coal-mining regions. But he has found something surprising about the towns and their inhabitants. His research shows that the DNA in these districts is flecked with disadvantage<sup>1</sup>, just as the coal seams once threaded through the ground.

By looking at the genomes of people living in former coal-mining areas, he has found genetic signatures associated with spending fewer years at school compared with people outside those areas, and – at weaker significance levels – variants that correlate with lower socio-economic status. Some genetic variants even correlate with political persuasion and whether or not communities voted to leave the European Union in the 2016 Brexit referendum.

Abdellaoui, who works at the University of Amsterdam in the Netherlands, acknowledges that he is venturing onto politically charged ground. "I try to understand human genetic

variation and this is what I run into," he says.

The study<sup>1</sup> – published this week in *Nature Human Behaviour* – is a high-profile example of an emerging trend: using huge amounts of data and computing power to uncover genetic contributions to complex social traits. Studies published in the past decade have examined genetic variants linked to aggression, same-sex sexual behaviour, well-being and antisocial behaviours, as well as the tendency to drink and smoke. In doing such science, geneticists are heading for controversial territory. They have even been accused of "opening a new door to eugenics", according to the title of a 2018 *MIT Technology Review* article by science historian Nathaniel Comfort<sup>2</sup>.

To the geneticists and social scientists doing this work, the results offer a useful and important guide to the relative contributions of nature and nurture to specific behavioural traits – just as genetic analysis can already highlight people who have an increased risk of cancer or heart disease. The approach could, for example, improve understanding of how the environment affects complex traits, and so offer a way to intervene to improve areas such as public education.

"It is super-exciting," says Philipp Koellinger, a geneoconomist at Vrije University Amsterdam in the Netherlands. "It gives us better and more-precise ways for scientists to answer questions they

have been interested in for a long time."

Caveats abound. The genetic contribution to any behavioural trait is relatively small and easily swamped by the influence of the environment. The studies can reveal only whether someone is likely to have a certain trait, and cannot predict the qualities of any one individual. Most scientists are quick to point out why they do this work – to establish what role, if any, genetics has in behaviour – and to lay out its limitations.

But not everyone is listening; already, some companies see a market in reading DNA like a fortune-teller reads tea leaves. "That stuff totally gives me the shivers. But it's happening," Koellinger says.

Critics charge that the ethical and societal risks of acting on such information are too great. "One of the main concerns is not so much the study of genomics, but how are we going to use it," says Maya Sabatello, a bioethicist at Columbia University in New York City. "Who's going to benefit? Who's not going to benefit? We live in a very unequal society and this is a major challenge."

## Strength in numbers

For decades, geneticists assumed that most traits were governed by just a handful of genes – whether it was a relatively simple one such as height, or something as complex as antisocial behaviour. But as the sample sizes swelled,



In former UK coal mining areas, genetic variants are linked with spending less time in school.

researchers began to find hundreds of variants that each have a relatively small effect on a trait. These projects – known as genome-wide association studies (GWAS) – build up a picture of which DNA letters vary from person to person (called single nucleotide polymorphisms, or SNPs), which variants are most common in people with a given trait and how much of the difference between individuals these SNP patterns represent.

Adding up the contributions made by all these spots on the genome gives researchers a measure of the importance of genetics to a trait, known as a polygenic score. For height, which is known to have a strong genetic influence, GWAS show that variants can together account for 20% of the variation.

As studies into physiology and disease piled up, scientists began to wonder whether the methods would work on social and psychological attributes.

For some complex traits, such as social isolation, researchers have found only a weak influence; one study<sup>3</sup> noted that heritability for that trait hovers at 4%. But for others, the signal from genetics studies has blossomed from initially feeble to surprisingly strong. In 2013, a large group of researchers working under the umbrella name The Social Science Genetic Association Consortium (SSGAC) reported the

first GWAS of educational attainment<sup>4</sup>, defined as years of schooling. The study found three SNPs that together could explain a meagre 2% of the variation in years of education. But then, a 2016 repeat by the same consortium using a sample that included almost 300,000 people – more than double the number in the 2013 study – found 74 SNPs that could explain 3.2% of the variation<sup>5</sup>. When the consortium combined data from 1.1 million people, they discovered more than 1,200 SNPs that together accounted for 11–13% of the variation<sup>6</sup>. That means the genes for educational attainment can explain about as much variation in a child's time in education as their family's socio-economic status can. "I think that's really quite remarkable," says Tim Morris, an epidemiologist at the University of Bristol, UK.

Beyond education, researchers have examined other socially shaped traits. In 2016, for instance, the SSGAC published a GWAS of almost 300,000 people and identified 3 SNPs associated with self-reported measures of well-being<sup>7</sup>. And in 2017, a weak genetic signature for antisocial behaviour showed up in a GWAS of a group of 6,200 Finnish prisoners<sup>8</sup>. Neither study produced a polygenic score, but researchers expect scores for these traits will emerge as sample sizes continue to grow.

The growing power of GWAS inspired

Abdellaoui to ask a different question: how do social traits such as educational attainment vary across a country? To find out, he and his team dug into the UK Biobank data set, which holds blood and tissue samples and survey responses for almost 450,000 people and cross-references the information to medical data such as hospital admissions.

The team looked at previous studies to amass a list of 33 health and behavioural traits and the genetic variants that influence them, adding up the contribution of each variant to get a polygenic score. The researchers then investigated the UK Biobank samples to see whether these genotypes differed across the United Kingdom. They first discounted genetic variation caused by historical regional differences in ancestry, throwing out variants that are common because of shared ancestry rather than because they govern a trait. Then they could see which traits still clustered into certain regions. For some traits – caffeine consumption, for example – there was no regional difference. But for others, such as educational attainment, the difference was significant. The researchers found that people living in former coal-mining regions had, on average, fewer genetic variants that correlated with staying in school longer or with going on to higher education<sup>1</sup>.

Peter Visscher, a geneticist at the University of Queensland in Australia who worked on the study, says it's not clear what underlying biology the genetic patterns identified represent. "I see that as a proxy for genes to do with intelligence and maybe perseverance, and maybe a bit of risk-taking."

Abdellaoui stresses that what they have produced is more description than explanation. "There are a whole bunch of variables that are clustering in the lower economic areas, but it's very difficult to say anything about directions of causality."

The researchers think the regional difference is down to the migration of more-educated people to richer areas that offer them jobs, leaving behind people who have genetic signatures linked to spending less time in school. This social stratification could become more marked over time, they say. "If that goes on for multiple generations, then for the sort of social inequalities already there, you run the risk of increasing those inequalities on a biological level," says Abdellaoui.

The researchers found the same geographic pattern for other traits, but the relationships were weaker. Genotypes known to be strongly associated with lower socio-economic status and lower cognitive ability were found more often in the poorer areas. These genotypes, the scientists reported, were associated with people's political views. Those in coal-mining areas had more genetic variants linked to lower socio-economic status, and were also more likely to vote for the left-wing Labour party or the right-wing UK Independence Party.



Individuals were also more likely to have voted for the United Kingdom to leave the EU in the Brexit referendum. Abdellaoui says this does not mean that someone is genetically predisposed to vote in a certain way.

Other researchers in the field agree with this caution. “Overall I like the paper and think that they’ve done a good job with it,” says Morris. “My main fear is that these results will be over-interpreted. They are informative descriptive statistics, but descriptive nonetheless.” He also notes the UK Biobank data are “extremely selective” and not likely to fully represent the populations of the former coal-mining regions. “For the regional results, these really must be interpreted with care.”

The results of this kind of study are based on associations, and must be presented very carefully to prevent suggestions that a person’s genes determine their outcomes, says Daniel Benjamin, a behavioural economist at the University of Southern California in Los Angeles. He is wary of comparisons between his field and the spectre of eugenics, an idea from the beginning of the twentieth century that people seen as having ‘inferior’ genes should be prevented from having children. “Those of us who do work in this area have an ethical obligation, and that ethical obligation is even stronger in the case of the genetics of behaviour because of past terrible misinterpretations and horrible consequences,” he says.

One of the biggest sources of confusion is what a polygenic score actually shows about the contributions of nature and nurture, Benjamin says. “People have a really hard time understanding that genes don’t determine behaviour.”

Abdellaoui says of his UK study: “We are in no way suggesting the genes are the sole determinant of someone’s educational outcome. It’s a combination of environmental and genetic effects.”

## Genetics in the classroom

Another disclaimer is that polygenic scores represent the ‘risk’ of having a particular trait, and don’t necessarily suggest that genetics is a major factor in behaviour. For instance, the scores cannot foretell that one individual will definitely graduate from university and another will quit school aged 16. “I don’t think that polygenic scores are at the level of predictive ability that would allow you to make those kinds of individual judgements with any degree of certainty,” says Paige Harden, a psychologist at the University of Texas at Austin.

When Benjamin and his team put together the most recent GWAS on education<sup>6</sup>, his team released an accompanying 20-page list of frequently asked questions to explain the study’s motives, which made clear that the scientists thought there were no implications for education policy. Not everyone is so cautious, says Morris. “There are quite a few academic papers

coming out that can’t resist a final sentence right at the end, along the lines of ‘the DNA revolution is coming and genes will soon be useful for predicting education’, which I think is quite irresponsible,” he says. He wants such papers to include more context – for example, pointing out that existing information such as a student’s previous attainment can already do a better job of predicting their future performance than a polygenic score can.

A working group announced earlier this month by bioethics think tank The Hastings Center in Garrison, New York, plans to examine the field and advise researchers and stakeholders on how to conduct and talk about the work (see [go.nature.com/2vtbpey](http://go.nature.com/2vtbpey)).

But others are less guarded. They argue that genetic screens of behaviour and cognitive

**“Many teachers are worried that trying to use genetics as a tool in education could potentially be misused.”**

ability could help children as young as three to fare better at school. “It can’t be right for education to continue to ignore genetic influence, because it’s far and away the most important source of individual differences,” says Robert Plomin, a psychologist at King’s College London, who is one of the more bullish voices in the debate and whose interpretations of the studies are controversial.

Sabatello, the bioethicist, predicts that the first applications will be in specialist education, such as for cases in which the parents of children with conditions such as attention deficit hyperactivity disorder (ADHD), autism spectrum disorder or dyslexia could use genotypes as evidence to demand a different approach for their child. “Parents want the genomic information to persuade authorities or educational entities that their kids need the specialist intervention.”

At the moment, there are no reliable polygenic scores to assess the contribution of genes to these conditions, but large-scale studies, more powerful than those done before, including a major GWAS currently under way for ADHD, could produce them in the future.

Although the focus on identifying and helping children with extra educational needs might sound altruistic, it, too, has a troubling historical precedent. Intelligence tests, which were first developed at the beginning of the twentieth century to pick out children who could benefit from extra attention, quickly became used to reinforce discrimination against minority populations or institutionalize children deemed to be ‘feeble-minded’.

“Many teachers are worried that trying to use genetics as a tool in education could potentially be misused to validate race and class-based

differences,” says Daphne Martschenko, who has just finished a PhD at the University of Cambridge, UK, that investigated attitudes in education to genetics.

In fact, because GWAS are done mostly using data from people of European ancestry, this could make the results less applicable for different ethnic groups. “A real pragmatic challenge is that we don’t have good genetic indicators for children of colour,” Harden says.

Morris thinks that this could compound existing inequality in education. “If you can’t do something for everyone in the system, then you can’t do it.”

## Responsible research

Many in the field agree that the most useful application of these results will be to allow better-quality research into environmental – not genetic – influences on complex behavioural traits, by taking out the influence of genetics while studying some other factor. “It’s an unsexy thing to talk about,” says Harden, “but a better idea is using genetics as a control variable to work out what actually works to improve learning.”

Researchers could include children with similar polygenic scores in both the control and test groups when trialling an intervention, for instance.

The results could also help scientists to probe whether the effects of genetics depend on an individual’s environment – whether certain gene variants kick in only under some circumstances. And more-sophisticated genetic studies could unpick the importance of something called genetic nurture, in which environmental influences are misidentified as genetic. This could be the case with education, because well-educated parents both pass on their genes and are more likely to contribute indirectly by encouraging their children’s schooling<sup>9</sup>.

The priority for most researchers in this field is to do more and bigger studies, to produce ever-stronger signals and tackle different traits such as income and social withdrawal. Meanwhile, those at the educational coalface don’t need insight from genetics to improve outcomes, says Sabatello. “We need to look at the environment. Children who are hungry can’t study. We don’t need to have their genes for that.”

**David Adam** is a freelance journalist based near London.

1. Abdellaoui, A. et al. *Nature Hum. Behav.* <https://doi.org/10.1038/s41562-019-0757-5> (2019).
2. Comfort, N. ‘Sociogenomics is opening a new door to eugenics’ *MIT Technol. Rev.* (23 October 2018).
3. Day, F. R., Ong, K. K. & Perry, J. R. B. *Nature Commun.* **9**, 2457 (2018).
4. Rietveld, C. A. et al. *Science* **340**, 1467–1471 (2013).
5. Okbay, A. et al. *Nature* **533**, 539–542 (2016).
6. Lee, J. J. et al. *Nature Genet.* **50**, 1112–1121 (2018).
7. Okbay, A. et al. *Nature Genet.* **48**, 624–633 (2016).
8. Tielbeek, J. J. et al. *JAMA Psychiatry* **74**, 1242–1250 (2017).
9. Kong, A. et al. *Science* **359**, 424–428 (2018).

# Books & arts

## Double deception in the asylum?

Susannah Cahalan's investigation of the social-psychology experiment that saw healthy people sent to mental hospitals finds inconsistencies. **By Alison Abbott**

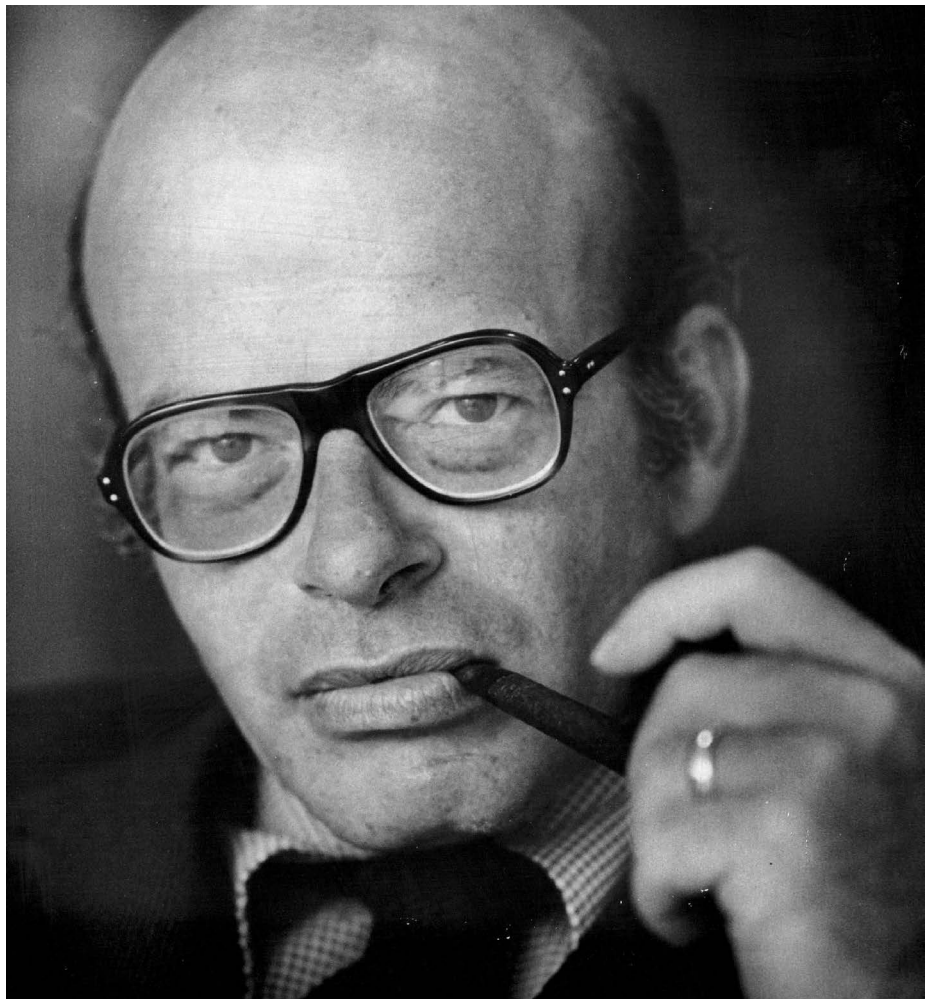
**F**rom 1969 to 1972, an extraordinary experiment played out in 12 psychiatric institutions across 5 US states. Eight healthy people – including David Rosenhan, a social psychologist at Stanford University in California, who ran the experiment – convinced psychiatrists that they needed to be committed to mental hospitals. The ensuing paper, published in *Science* in 1973, opens with the words: “If sanity and insanity exist, how shall we know them?” It claimed that the psychiatric establishment was unable to distinguish between the two.

Rosenhan's study had far-reaching and much-needed effects on psychiatric care in the United States and elsewhere. By the 1980s, most psychology textbooks were quoting it. It also influenced society more widely, and not always positively: in the law courts, for instance, it undermined the value of expert testimonies from psychiatrists. Now, in *The Great Pretender*, journalist Susannah Cahalan turns a fresh, critical eye on the experiment and the shockwaves it sent through the field and beyond.

Cahalan quotes a former colleague of Rosenhan's, who notes that he was a good networker, an excellent lecturer and a generally charismatic character. “But some people in the department called him a bullshitter,” Kenneth Gergen says. And through her deeply researched study, Cahalan seems inclined to agree with them. She discovered that the man whom she had initially admired, and who had done so much to change how mental illness was perceived, was not all that he had seemed. And neither, she argues, was his famous experiment.

Cahalan began her investigation into Rosenhan's experiment in good faith. Ten years ago, she developed paranoia, hallucinations and, eventually, seizures. She was dosed with antipsychotics before being correctly diagnosed with a very rare type of autoimmune encephalitis, an ordeal she describes in her first book, *Brain On Fire*.

After it was published in 2012, a casual conversation with McLean Hospital psychiatrists in Boston, Massachusetts, alerted Cahalan



David Rosenhan and his volunteers feigned symptoms to be admitted to psychiatric hospitals.

to Rosenhan's experiment. She immediately wanted to know more – about the experiences of those who volunteered, and the challenges that such a risk-laden experiment would have posed decades ago.

Rosenhan was not the first to infiltrate a psychiatric hospital and report on conditions. Cahalan tells, for example, of the nineteenth-century journalist Nellie Bly, who deceived doctors to spend ten days in an overcrowded women's asylum on Blackwell's Island, New York. Bly's reports of the appalling

conditions there shamed politicians into increasing the asylum's budget.

But Rosenhan was the first to carry out a formal experiment involving a number of “pseudopatients”. All eight, including Rosenhan, reported the same symptom to different doctors: that they heard voices uttering “thud, empty, hollow”, denoting existential doom. Seven were diagnosed with schizophrenia; one with manic depression. Once admitted to hospital, the volunteers stopped simulating symptoms of abnormality. Rosenhan noted

in the *Science* paper that genuine patients often realized that the pseudopatients did not have a mental-health disorder, and accused them of being undercover journalists or academics checking up on the hospital. Psychiatrists seemed less perceptive: it was several weeks before some of the pseudopatients got discharged.

Although Rosenhan died in 2012, Cahalan easily tracked down his archives, held by social psychologist Lee Ross, his friend and colleague at Stanford. They included the first 200 pages of Rosenhan's unfinished draft of a book about the experiment.

At first, it seemed that Cahalan's research was going to be easy, even though Rosenhan had given fictitious names to the pseudopatients she wished to track down, along with the hospitals they went to. Ross warned her that Rosenhan had been secretive. As her attempts to identify the pseudonymous pseudopatients hit one dead end after the other, she realized Ross's prescience.

The archives did allow Cahalan to piece together the beginnings of the experiment in 1969, when Rosenhan was teaching psychology at Swarthmore College in Pennsylvania. The students complained that the course was too abstract, so Rosenhan suggested that they check into a psychiatric hospital to get to know people with schizophrenia personally. The superintendent of the local Haverford State Hospital was willing to take them on, but Rosenhan cautiously decided to check things out for himself first. He emerged humbled from nine traumatizing days in a locked ward, and abandoned the idea of putting students through the experience. But it set him thinking about a scientific experiment aimed at exposing the system's travesties.

According to Rosenhan's draft, it was at a conference dinner that he met his first recruits: a recently retired psychiatrist and his psychologist wife. The psychiatrist's sister also signed up. But the draft didn't explain how, when and why subsequent recruits signed up.

Cahalan interviewed numerous people who had known Rosenhan personally or indirectly. She also chased down the medical records of

individuals whom she suspected could have been involved in the experiment, and spoke with their families and friends. But her sleuthing brought her to only one participant, a former Stanford graduate student called Bill Underwood.

**"Patients often realized that the volunteers were not ill, and accused them of being undercover journalists."**

Underwood and his wife were happy to talk, but two of their comments jarred. Rosenhan's draft described how he prepared his volunteers very carefully, over weeks. Underwood, however, remembered only brief guidance on how to avoid swallowing medication by hiding pills in his cheek. His wife recalled Rosenhan telling her that he had prepared writs of habeas corpus for each pseudopatient, in case an institution would not discharge them. But Cahalan had already worked out that that wasn't so.

Comparing the *Science* report with documents in Rosenhan's archives, she also noted many mismatches in numbers. For instance,

Rosenhan's draft, and the *Science* paper, stated that Underwood had spent seven days in a hospital with 8,000 patients, whereas he spent eight days in a hospital with 1,500 patients.

When all of the leads from her contacts led to ground, she published a commentary in *The Lancet Psychiatry* asking for help in finding them – to no avail. Had Rosenhan invented them, she found herself asking?

In recent years, other heroes of social psychology have been found to have misrepresented their data. The most prominent case is that of Dutch social psychologist Diederik Stapel, who was forced to retract 58 papers. Those who have followed these cases might be appalled by the Rosenhan story, but will not be surprised.

Cahalan, whose life was saved by front-line medical science in the context of psychiatry, was shocked by what she found. She writes that she cannot be completely certain that Rosenhan cheated. But she is confident enough to call her engrossing, dismaying book *The Great Pretender*.

**Alison Abbott** writes from Munich, Germany.  
e-mail: [alison.abbott.consultant@springernature.com](mailto:alison.abbott.consultant@springernature.com)

## The rise of the greedy-brained ape

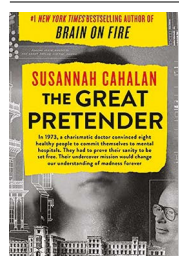
Gaia Vince takes an enjoyable sprint through human evolution. **By Tim Radford**

**G**aze into a mirror. Reflected is a marvel of evolution: a weak-jawed, bipedal omnivore with a greedy brain, in which 100 billion neurons consume 20% of the body's energy intake. Science journalist Gaia Vince urges us towards such reflections in *Transcendence*, a book tracing the journey of *Homo sapiens* through genes, environment and culture to what might be, she surmises, a new state of being.

For her hugely enjoyable sprint through human evolutionary history, Vince (erstwhile news editor of this journal) intertwines many threads: language and writing; the command of tools, pursuit of beauty and appetite for trinkets; and the urge to build things, awareness

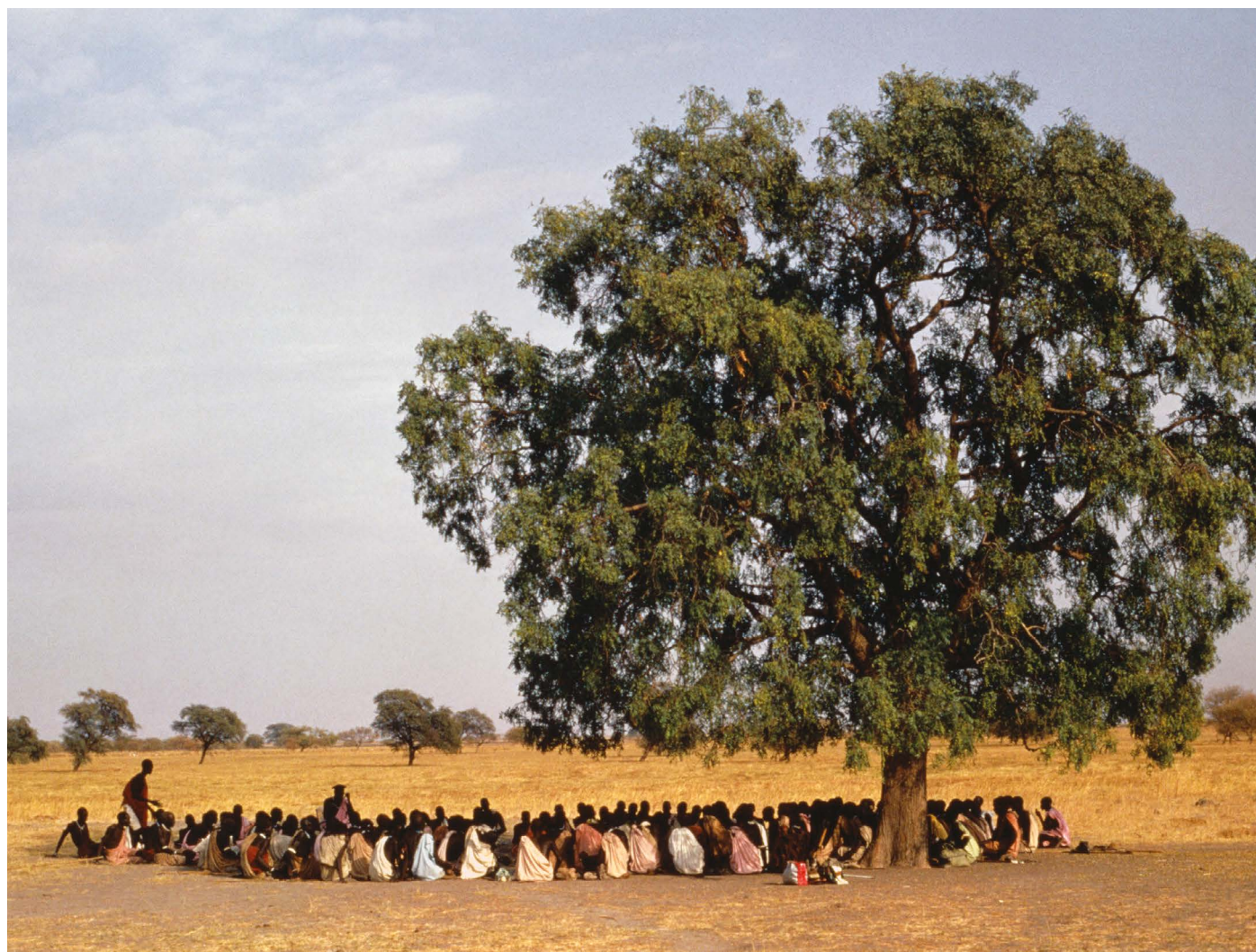
of time and pursuit of reason. She tracks the cultural explosion, triggered by technological discovery, that gathered pace with the first trade in obsidian blades in East Africa at least 320,000 years ago. That has climaxed this century with the capacity to exploit 40% of the planet's total primary production.

How did we do it? Vince examines, for instance, our access to and use of energy. Other primates must chew for five hours a day to survive. Humans do so for no more than an hour. We are active 16 hours a day, a tranche during which other mammals sleep. We learn by blind variation and selective retention. Vince proposes that our ancestors enhanced that process of learning from each other with the command



**The Great Pretender: The Undercover Mission that Changed our Understanding of Madness**  
Susannah Cahalan  
Grand Central Publishing (2019)





EYE UBIQUITOUS/ALAMY

Shilluk people in Sudan gather in the shade for traditional storytelling.

of fire: it is 10 times more efficient to eat cooked meat than raw, and heat releases 50% of all the carbohydrates in cereals and tubers.

Thus *Homo sapiens* secured survival and achieved dominance by exploiting extra energy. The roughly 2,000 calories ideally consumed by one human each day generates about 90 watts: enough energy for one incandescent light bulb. At the flick of a switch or turn of a key, the average human now has access to roughly 2,300 watts of energy from the hardware that powers our lives – and the richest have much more.

Humans are more social than other primates. We can keep track of around 150 other people, which demands a large brain and might also

help to expand it. To learn a fact stimulates one part of the brain; to hear a story activates many. That is why we find information 22 times more memorable in narrative form. *Homo sapiens* is a storytelling animal, and this adaptation ensures the transmission of skills and knowledge as fable, epic or cautionary tale. Vince, drawing on brain-scan studies, shows that neuroscientists have noted a synchrony, both spatial and temporal, between speaker and listener during storytelling, a phenomenon known as ‘neural coupling’.

The human capacity for narrative, metaphor and pattern-matching can lead us to see meaning where there is none, however. In a US psychological experiment in 1944, students were shown a short animation of two triangles and a circle passing across a screen, while a rectangle remained stationary (F. Heider & M. Simmel *Am. J. Psychol.* 57, 243–259; 1944). Of the subjects, 33 out of 34 anthropomorphized the moving shapes, creating narratives of anxiety, concern, rage and frustration.

Vince continually returns to the evolutionary triad of genetics, environment and culture to address our similarities and differences.

Some human biological adaptations are part of cultural variety. The semi-nomadic Moken people of Thailand can see clearly underwater because they can constrict their pupils to the maximum limit of human capacity, increas-

**“We find information 22 times more memorable in narrative form.”**

ing depth of field and changing the lens shape. This is a learnt capacity: in an experiment, Swedish children mastered it. Divers of the Bajau people in Indonesia, however, exemplify heritability and environmental selection at work. Their spleens are 50% larger than average, acting as a reservoir of oxygenated blood and endowing them with consummate endurance underwater.

Our most profound cultural tool, language, is in some ways culturally selected. We owe our acrobatic way with words to a larynx that descends at three months of age. Thereafter, Vince notes, we can no longer swallow and



**Transcendence:**  
**How Humans**  
**Evolved through Fire,**  
**Language, Beauty,**  
**and Time**  
Gaia Vince, Allen Lane  
(2019)



breathe at the same time. Our languages shape our thinking and cultural identity in many ways, but environment also shapes speech. Languages in warm, wet, wooded regions tend to have more vowels and fewer consonants. Languages that emerged at altitude have more words with a strong expulsion of air in the consonants.

Tonality in languages (in which a word has different tones that change meaning) is important. The emergence of non-tonal languages over the past 50,000 years – Homer’s Greek was tonal, modern Greek is not – might have influenced the spread of two gene variants involved in brain growth, according to a 2007 study (D. Dediu and D. R. Ladd *Proc. Natl Acad. Sci. USA* **104**, 10944–10949; 2007). So words also shape our inheritance.

Vince has a lot to say about words. The average response rate between speakers during a conversation is 200 milliseconds. But it takes 600 milliseconds for the signal to go from ears to brain, to understanding, to the preparation of a response and its transmission. Thus, conversation must rely on a sophisticated prediction system that commits a large part of the brain to both speaking and listening. Language, writes Vince, “gives us an unparalleled ability to convey an infinity of ideas. We use it mainly to talk about ourselves.”

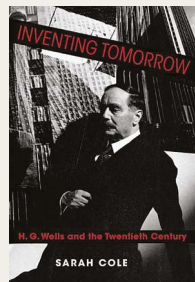
Of course we do. Humans might not be so much *Homo sapiens* as *Homo narcissus*, the self-absorbed species. Yet all of our capacities together have, in their different ways, endowed us with the capacity to become a super-organism. We are now a globally connected urban species, outsourcing our brains to computers, increasingly to artificial intelligence and (so far) to nine billion robots. We have begun the Anthropocene, and our demands on the planet are not sustainable. That could usher in a new dark age, or a global order in a new shared civilization. We transcend our evolutionary beginnings.

Vince dubs this emerging species *Homo omnis*, or *Homni* for short. Her chosen analogue for such a biological super-organism is not flattering: it is the slime mould, in which single cells coalesce as one to move on. The fortunate are protected at the centre; those on the margin become vulnerable to environmental change. Which sounds disturbingly like us.

Many aspects of *Transcendence* have been explored before. And, with that wealth of palaeoanthropological and other research to draw from, most of the chapters become a mosaic of tersely introduced evidence. Read it anyway. It is at least 22 times more memorable than many textbooks, and a good story without – so far – a happy ending.

**Tim Radford** is a former science editor of *The Guardian*. His book *The Consolations of Physics* is published by Sceptre.  
e-mail: radford.tim@gmail.com

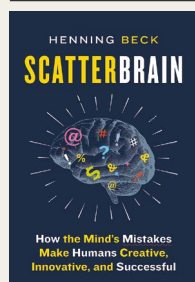
## Books in brief



### Inventing Tomorrow

Sarah Cole Columbia University Press (2019)

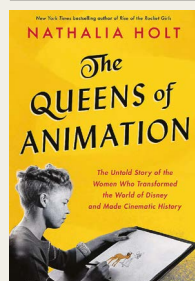
H. G. Wells was, asserts scholar Sarah Cole, a pioneer adept at “rescaling the cosmos and humanity’s place in it”. He straddled the border between science and literature, but not all his complexities were benign: he both repudiated racism and for some time shamefully ascribed to ideas on eugenics. Cole adroitly captures Wells, from his mould-breaking books (such as the 1895 science-fiction classic *The Time Machine* and 1920 *Outline of History*) to his unlikely intellectual kinship with subtle modernists such as Virginia Woolf.



### Scatterbrain

Henning Beck Greystone (2019)

The human brain at work, notes neuroscientist Henning Beck, is sloppy – and that is precisely what makes us creative powerhouses. Beck’s coolly amusing narrative takes us through forgetting, pigeonholing, distraction and deep into creativity. He explores how idle wool-gathering is more conducive to creativity than is ‘efficient’ thinking, and the uncannily similar way in which true and false memories are generated in the brain. His is a hopeful message, ultimately. If we don’t err, we don’t change. So: “stay fallible”.



### The Queens of Animation

Nathalia Holt Little, Brown (2019)

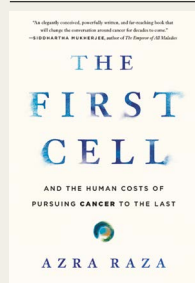
The early hand-drawn animations of Walt Disney Studios remain a technological wonder. Few know, however, of the company’s female virtuosi, who from the 1930s on injected nuance into characters from Bambi to a panoply of princesses. In her gripping corrective, Nathalia Holt ushers these animators and story developers into the limelight: Bianca Majolie, Sylvia Holland, Retta Scott, Grace Huntington and Mary Blair. Particularly in the early years, Holt shows, they paid a high price to work, forced to battle harassment in mostly male teams.



### Rebel Star

Colin Stuart Michael O’Mara (2019)

This compelling portrait of the Sun packs in facts while speculating on gaps in our knowledge. Astronomy journalist Colin Stuart traces the arc of discovery from the fourth-century BC heliocentricism of Aristarchus of Samos through solar spectroscopy, star formation and nuclear fusion, the “epic journey” of sunlight to Earth and more. The Sun is both bountiful and belligerent, he reminds. Solar power could make 87% of countries energy self-sufficient – but the next big solar storm could send our electrical infrastructure into meltdown.



### The First Cell

Azra Raza Basic Books (2019)

Each year, the United States spends US\$150 billion on treating cancer. Yet as oncologist Azra Raza notes in this incisive critique-cum-memoir, the treatments remain largely the same. Raza wants to see change: eliminating the first cancer cell rather than “chasing after the last”, which is doable with current technologies. Meanwhile, she braids often-harrowing stories of patients, including her own husband, with insights gleaned from laboratory and literature on this complex, often confounding array of diseases. **Barbara Kiser**

# Comment



ILLUSTRATION BY SEÑOR SALME

## Ethical research – the long and bumpy road from shirked to shared

Sarah Franklin

From all too scarce, to professionalized, the ethics of research is now everybody's business, argues Sarah Franklin in the sixth essay in a series on how the past 150 years have shaped science.



**Anniversary collection:**  
[go.nature.com/nature150](https://go.nature.com/nature150)

In the autumn of 1869, Charles Darwin was hard at work revising the fifth edition of *On The Origin of Species* and drafting his next book, *The Descent of Man*, to be published in 1871. As he finished chapters, Darwin sent them to his daughter, Henrietta, to edit – hoping she could help to head off the hostile responses to his debut, including objections to the implication that morality and ethics could have no basis in nature, because nature had no purpose.

That same year, Darwin's cousin Francis Galton published *Hereditary Genius*, a book that recast natural selection as a question of social planning<sup>1</sup>. Galton argued that human abilities were differentially inherited, and introduced a statistical methodology to aid “improvement of the race”. Later, he coined the term ‘eugenics’

to advocate selective reproduction through application of the breeder's guiding hand.

Darwin's transformative theory inspired modern biology; Galton's attempt to equate selection and social reform spawned eugenics. The ethical dilemmas engendered by these two late-nineteenth-century visions of biological control proliferate still. And, as older quandaries die out, they are replaced by more vigorous descendants. That there has never been a border between ethics and biology remains as apparent today as it was 150 years ago. The difference is that many of the issues, such as the remodelling of future generations or the surveillance of personal data, have become as everyday as they are vast in their implications. To work out how to move forward, it is worth looking at how we got here.





WILL LATHAM/EYEVINE

Embryologist Ian Wilmut encounters his brainchild Dolly the cloned sheep, which is now stuffed and went on show at an exhibition in 2015.

In the late nineteenth century, like today, society was in upheaval and science was on a roll. With Darwin's bold hypotheses set before them, Victorian breeders, microscopists, collectors, astronomers, geologists and anatomists sought to discover the laws interconnecting life's core processes – often by using ingenious experimental designs. To probe the formative effects of gestation on heredity in mammals, the gentleman naturalist Walter Heap, a laboratory demonstrator at the University of Cambridge, UK, conducted the first experiments in transferring embryos from one variety of rabbit to another at his home in Prestwich in the 1890s. His methods typified a new era of disrupt-and-learn biology.

### Biology rebooted

By the early part of the twentieth century, what had come to dominate was “the biological gaze”, to quote historian Evelyn Fox Keller at the Massachusetts Institute of Technology in Cambridge<sup>2</sup>. Rather than simply observing life, experimenters began to manipulate its component parts to test the limits of the system, mix up ingredients and turn biology inside out.

In 1903, the embryologist Hans Spemann conducted his famous experiments with amphibians. Using one of his infant daughter's fine, elastic hairs, he tied a loop around a fertilized salamander egg to create an animal

with two heads and one tail. That same decade, in the United States, physiologist Jacques Loeb pursued a new ‘engineering biology’, trying out all sort of chemicals and conditions to prompt development in model organisms such as sea urchins<sup>3</sup>.

**“The bar for proper scrutiny has not so much been lowered as sawn to pieces.”**

Ian Wilmut, who led the team that created Dolly the cloned sheep in the 1990s (at the Roslin Institute near Edinburgh, UK), once stated that it was Dolly's birth that ushered in “the age of biological control” – and made obsolete the expression “biologically impossible”<sup>4</sup>. In fact, this view of life was born at least a century earlier. And as confident experimentation turned ever more closely and deliberately towards humans, the relationships between research, industry and governments became a tangled ethical bank, and have remained so ever since.

Eugenics, never without its trenchant opponents, became an increasingly crucial part of a new world order over the course of the twentieth century. It is particularly associated with the mass-sterilization campaigns that began after Indiana's 1907 act, and with the Nazi

racial-hygiene programme that reached its nadir in the Holocaust.

Another legacy of the eugenics movement is the management of populations using techniques such as demography, racial classification and statistical modelling. These, combined with family planning, became synonymous with modernity and progress. From Latin America and Scandinavia to India, China and the Soviet Union, eugenics took root in projects to ‘improve the population’ throughout the twentieth century. Eugenic presumptions about the differential fitness of native and immigrant populations were central to colonial administrations across the British Empire. Census-takers created ‘races’ and ‘tribes’ where none existed, for the purpose of managing populations more ‘scientifically’. These categorizations got inked into emerging nations across Africa and southeast Asia, and continue to shape definitions of race in countries including Malaysia and Singapore today.

The logic of the modern nation state is in no small part provided by eugenic techniques of classifying and controlling citizens, as pointed out by historians Alison Bashford, now at the University of New South Wales in Sydney, Australia, and Philippa Levine, at the University of Texas at Austin<sup>5</sup>. This typological approach to administration was normalized through what has been called the “prism of heritability” by the

sociologist Troy Duster, now at the University of California, Berkeley<sup>6</sup>. That had the effect of linking together the pathologization of mental illness, homosexuality, criminality, poverty, ethnicity and race into a discourse of ‘rational’ management that became mainstream.

In other words, the principles of the eugenics movement are part of contemporary society’s DNA. Across national and global policies affecting everything from health care, fertility and incarceration to border control, education and regional development, the goal of shaping the population through selective pressures – such as creating a “hostile environment” for immigrants – is alive and well.

## The rise of bioethics

The birth of bioethics in the 1970s was in no small part a response to harmful research projects undertaken within this context – on vulnerable groups such as immigrants, prisoners and psychiatric patients – and without meaningful consent. The field emerged largely in the United States, partly driven by the international outrage at the exposure in 1972 of the covert US Public Health Service research project at Tuskegee University – in which more than 400 black US men, mostly poor share-croppers from Alabama, were subjected to untreated syphilis between 1932 and 1972. As many as half of them died, and 60 of their wives and children contracted the disease.

In 1974, the US government passed the National Research Act and established the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. Two years later, the commission drafted a report outlining in detail the “basic ethical principles and guidelines that should assist in resolving the ethical problems that surround the conduct of research with human subjects”.

In 1978, this was published as *The Belmont Report*<sup>7</sup> in the US Federal Register, establishing guidance for national research and the three pillars of modern bioethics. These were: respect for persons, beneficence (‘doing good’) and justice. The report also clarified the basis for informed consent of study participants, and helped to enforce mandatory policies for ethical oversight of research. The three principles were largely aimed at preventing the mistreatment of vulnerable individuals and communities. Under Belmont’s influence, research ethics became a central principle of modern science.

Bioethics flourished throughout the 1980s, expanding to include equity in public health and access to medical care. The field became increasingly central to medical and scientific training, as well as to research funding. That focus was intensified by the ‘too little, too late’ critiques of government responses to the HIV crisis that emerged in the mid-1980s.

Bioethics gathered momentum at this time by offering guidance on controversial

biomedical applications such as organ transplants and *in vitro* fertilization (IVF). In the first encyclopaedia of bioethics, published in 1978, theologian Warren T. Reich drew attention to a key shift in the practice of medicine: one that moved away from a commitment to preserving life<sup>8</sup>. In the past, he argued, medicine was guided by the absolute principle to ‘do no harm’. However, a different ethical dilemma arose out of a heart transplant, a procedure that could significantly improve an individual’s quality of life but which also had the potential to kill them. In contrast to the iron-clad medical ethics of old, the absolute value of human life became relativized. In the world’s most advanced medical facilities, a higher quality of life could now be worth dying for. Once again, ethical debate was reignited.

It was in the 1990s that professionalized bioethics reached its high-water mark. The Human Genome Project (HGP) – the leviathan of publicly funded DNA sequencing – promised to unleash a combination of Darwin’s and Galton’s visions as the century drew to a close. Ethics claimed the largest share of HGP funds set aside for the analysis of “Ethical, Legal, and Social Implications” (ELSI) of genome mapping.

## “Bioethics, once a beacon of principled pathways to policy, is increasingly lost.”

In the United States alone, around US\$3 billion (5% of the HGP budget) was spent to create “the world’s largest bioethics program”. Armies of ethicists combed over the philosophical principles of altering genetic material in ways that might or might not be passed on to future generations, and the perils of designer babies.

Then, as the century neared its end, something else took centre stage: new techniques derived from reproductive and developmental biology, such as cloning and research into stem cells and embryos. As the prospects of quick bench-to-bedside applications from the HGP faded, so did the allure of bioethics. The discipline lost its most significant source of funding as ELSI programmes ceased.

To return to 1978, there was another turning point for bioethics in the year of *The Belmont Report*: the birth of Louise Brown, the first baby conceived through IVF. Some of the most controversial research and applications over the past half-century have concerned reproductive and developmental biology. But while bioethicists were recruited en masse to contemplate the impact of the HGP, the fertility industry mushroomed, generating an impressive set of acronyms (but no ELSI). For a while, global public opinion became more sharply divided over cloned dogs and genetically modified (GM) maize (corn) than GM babies. That concern would come later.

In retrospect, many of the forces that propelled late-twentieth-century bioethics into the limelight – such as the focus on speculative genomic futures – eventually left it unmoored. In the past two decades, bioethics has drifted into uncharted waters. Today, amid a panoply of ethical quagmires ranging from gene-edited babies and neurotechnology to dish-grown organoids and nanobots, the fraught relationship between society and research is once again front and centre.

## Beyond bewilderment

Just as the ramifications of the birth of modern biology were hard to delineate in the late nineteenth century, so there is a sense of ethical bewilderment today. The feeling of being overwhelmed is exacerbated by a lack of regulatory infrastructure or adequate policy precedents. Bioethics, once a beacon of principled pathways to policy, is increasingly lost, like Simba, in a sea of thundering wildebeest. Many of the ethical challenges arising from today’s turbocharged research culture involve rapidly evolving fields that are pursued by globally competitive projects and teams, spanning disparate national regulatory systems and cultural norms. The unknown unknowns grow by the day.

The bar for proper scrutiny has not so much been lowered as sawn to pieces: dispersed, area-specific ethical oversight now exists in a range of forms for every acronym from AI (artificial intelligence) to GM organisms. A single, Belmont-style umbrella no longer seems likely, or even feasible. Much basic science is privately funded and therefore secretive. And the mergers between machine learning and biological synthesis raise additional concerns. Instances of enduring and successful international regulation are rare. The stereotype of bureaucratic, box-ticking ethical compliance is no longer fit for purpose in a world of CRISPR twins, synthetic neurons and self-driving cars.

Bioethics evolves, as does any other branch of knowledge. The post-millennial trend has been to become more global, less canonical and more reflexive. The field no longer relies on philosophically derived mandates codified into textbook formulas. Instead, it functions as a dashboard of pragmatic instruments, and is less expert-driven, more interdisciplinary, less multipurpose and more bespoke. In the wake of the ‘turn to dialogue’ in science, bioethics often looks more like public engagement – and vice versa. Policymakers, polling companies and government quangos tasked with organizing ethical consultations on questions such as mitochondrial donation (‘three-parent embryos’, as the media would have it) now perform the evaluations formerly assigned to bioethicists. Journal editors, funding bodies, grant-review boards and policymakers are increasingly the new ethical adjudicators.

These shifts have been a long time coming



## Comment

and have many different sources, including the driving influence of practical ethicists such as the British philosopher Mary Warnock – often to the consternation of the wider bioethics community. After Warnock's *Report of the Committee of Inquiry into Human Fertilisation and Embryology* was published for the UK government in 1984, John Harris, a medical ethicist at the University of Manchester, UK, complained that “the crucial questions are fudged, or rather are never addressed”<sup>9</sup>. He argued that Warnock's approach was over-reliant on “primitive feelings”, resulting in recommendations that were “false” and “dangerous”. The Warnock committee, in his view, had evaded the single most important question they faced concerning the moral status of the human embryo, in favour of a sentimental concession to expedient policy.

As we now know, Warnock was prescient in her attention to the strength of public feeling in relation to human-embryo research. Her reliance on several overlapping types of argument to justify strict limits on the introduction of new reproductive technologies has enabled the United Kingdom to establish a licensing system that is more flexible and that has proved more long-lasting than in any other country. Her committee, unusually comprising a majority of non-scientists, reached its consensus based on a pragmatic and principled proposal: that approval for the study of controversial therapeutic and experimental procedures would be subject to a strict and comprehensive code of practice upheld by Parliament. The law itself, Warnock argued, would act as both a guarantor and a symbol of public morality; it would in its combination of permissive scope and legislative precision express “the moral idea of society”. This was a

new template for ethical reasoning.

When the UK government decided, in the wake of the passage of the Human Fertilisation and Embryology Act in 1990, that it would not establish a counterpart to the US National Bioethics Advisory Commission, it was following Warnock's lead, and that of Anne McLaren. This developmental biologist and Warnock committee member took a populist and practical approach to public trust in science that has been highly influential.

Today, interdisciplinary expertise plus extensive and creative public consultation increasingly define a new approach to ethical science. This trend has been reinforced by organizations such as the Nuffield Council on

**“The most ethical science is the most sociable one... scientific excellence depends on greater inclusivity.”**

Bioethics, which advises the UK government by mobilizing a broad spectrum of knowledge, far beyond that of bioethicists and philosophers. Since 1993, the council has commissioned and published nearly 30 specialist reports on controversial biomedical issues, ranging from genetic screening to xenotransplantation. Few of the panels have been chaired by bioethicists. Many of the reports have widened the idea of what counts as an ethical issue – for example, the exploration of cultures of UK scientific research, chaired by the University of Cambridge plant scientist Ottoline Leyser<sup>10</sup>. In a similar vein, the International Society for Stem Cell Research has released a series of global guidelines that prioritize oversight,

communication and research integrity as well as patient welfare, social justice and respect for study participants over fixed principles of ethical conduct.

In a social-media-saturated age wary of fake news, the new holy grail is the ability to create trustworthy systems for governing controversial research such as chimeric embryos and face-recognition algorithms. The pursuit of a more ethical science has come to be associated with building trust by creating transparent processes, inclusive participation and openness to uncertainty, as opposed to distinguishing between ‘is’ and ‘ought’.

In short, expert knowledge and reliable data are essential but never enough to enable enduring, humane governance to emerge. So there is now more emphasis on continuous communication and outreach, and on long-term strategies to ensure collective participation and feedback at all stages of scientific inquiry. The result is less reliance on specialized ethical expertise and more attention to diversity of representation.

Amid the perils and promises of applications, from replacement heart and liver cells or driving malarial resistance through the mosquito population to ending Huntingdon's disease, a new legacy to Darwin and Galton has emerged. It turns out that what we have in common is less a single biological essence – or the ability to alter it – than a shared responsibility for human and non-human futures. The implication of this new model is that the most ethical science is the most sociable one, and thus that scientific excellence depends on greater inclusivity. We are better together – we must all be ethicists now.

## The author

**Sarah Franklin** is chair of sociology and director of the Reproductive Sociology Research Group at the University of Cambridge, UK.

e-mail: sbf25@cam.ac.uk

1. Galton, F. *Hereditary Genius: An Inquiry into its Laws and Consequences* (Macmillan, 1869).
2. Keller, E. F. in *FutureNatural: Nature, Science, Culture* (eds Robertson, G. et al.) 107–121 (Routledge, 1996).
3. Pauly, P. J. *Controlling Life: Jacques Loeb and the Engineering Ideal in Biology* (Oxford Univ. Press, 1987).
4. Wilmut, I., Campbell, K. & Tudge, C. *The Second Creation: The Age of Biological Control by the Scientists who Cloned Dolly* (Headline, 2000).
5. Bashford, A. & Levine, P. (eds.) *The Oxford Handbook of the History of Eugenics* (Oxford Univ. Press, 2010).
6. Duster, T. *Backdoor to Eugenics* (Routledge, 1990).
7. Ryan, K. J. et al. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research* (US Department of Health, Education and Welfare, 1978).
8. Reich, W. T. (ed) *Encyclopedia of Bioethics* Vol. 1 (The Free Press, 1978).
9. Harris, J. *The Value of Life: An Introduction to Medical Ethics* (Routledge & Kegan Paul, 1985).
10. Leyser, O. *The Culture of Scientific Research in the UK* (Nuffield Council on Bioethics, 2014).



**Louise Brown, the world's first baby born as a result of *in vitro* fertilization, pictured in 1981.**



Pregnant women in Lima, Peru, painted each other in 2013 to help a local hospital to raise awareness of good maternity health care.

# Prevent depression in pregnancy to boost all mental health

Ricardo F. Muñoz

**Tackling mental disorders before they arise in pregnant women and new mothers is an approach that could be scaled up online – and would aid the overall health of populations.**

I have been convinced of the importance of prevention in addressing mental-health problems since the early 1970s, when I began my doctorate in clinical psychology. But only now is there sufficient evidence from clinical trials of the effectiveness of preventive interventions, using approaches derived from interpersonal and cognitive behavioural therapy, to justify deploying them. And only now are the tools available to make such interventions available to people worldwide.

Two recent reports underline this conclusion. In February, the US Preventive Services Task Force, an independent panel of

experts in evidence-based medicine, urged clinicians to “provide or refer pregnant and postpartum persons who are at increased risk of perinatal depression to counseling interventions”<sup>1</sup>. And last month, the US National Academies of Sciences, Engineering, and Medicine (NASEM) released a report<sup>2</sup> calling on various stakeholders, from educators to policymakers, to prevent mental-health disorders and to promote healthy mental, emotional and behavioural development in the under 25s. (I was a member of the committees that prepared this document and two previous NASEM reports in 1994 and 2009 on preventive interventions<sup>3,4</sup>.)

The latest NASEM call to action<sup>2</sup> is so all-encompassing, it is hard to know where to begin. I propose that initial efforts focus on preventing depression in pregnant women or in women who have recently given birth (perinatal depression). There is substantial evidence for the effectiveness of providing such women with basic skills in mood management<sup>5</sup>. These interventions could have an impact across generations, because better

maternal mental health is linked to babies’ healthier development<sup>2</sup>. And if researchers and health-care systems were to monitor and compare the epidemiology of depression in thousands of mothers and their children in areas that have or have not deployed preventive interventions, stakeholders could measure their effect on entire communities.

Ultimately, massive open online interventions will need to be created (similar to the massive open online courses that are delivered on the Internet for free). These would allow anyone to obtain information and tools to help them stave off depression, at times and places that are convenient to them.

## A global challenge

In the United States, nearly 15% of men and around 26% of women experience a major depressive episode at some point<sup>6</sup>. People are diagnosed with this if they report experiencing five of nine symptoms over at least two weeks. These must include either feeling depressed or being unable to feel interest or pleasure, as



## Comment

well as problems sleeping, changes in appetite, fatigue or having suicidal thoughts.

Numerous psychological, pharmacological and physical treatments are effective, such as cognitive behavioural therapy, antidepressant drugs and electroconvulsive therapy. But many people who are depressed are not receiving treatment<sup>7</sup> because they fear stigma, can't get to clinics or afford treatment, or because there aren't enough psychologists and psychiatrists to meet their needs.

Given these challenges – and especially given the scale of the problem – societies worldwide need to take steps to stop depression from taking hold in the first place.

The number of randomized controlled trials testing preventive interventions has greatly increased since 1995 (see 'Mounting data'). Two approaches have been studied the most: cognitive behavioural therapy and interpersonal therapy. The first involves teaching people how to use the natural relationship between thoughts, behaviours and mood to increase those thoughts and behaviours that lead to healthy mood states – and to reduce or modify those that elicit sadness, helplessness and hopelessness. (People might be asked, for instance, to predict how their mood would change if they undertook certain activities, such as seeing a friend – and then to record how their mood actually changed following the activity.) The second approach, interpersonal therapy, helps people to communicate better with others, and so to obtain more support from friends and family.

In the early 2000s in California, for example, my colleagues and I at San Francisco General Hospital (now the Zuckerberg San Francisco General Hospital and Trauma Center) conducted a pilot study funded by a US National Institute of Mental Health grant. The study involved 41 Spanish- and English-speaking women, most of whom were in their 16th week of pregnancy. These women were not clinically depressed but were deemed to be at high risk because they scored 16 or more on a depression scale, or had a history of major depressive episodes. The preventive intervention we used in this case involved psychologists teaching a cognitive behavioural 'Mothers and Babies/Mamás y Bebés' course in 2-hour sessions once a week for 12 weeks. Only 14% of the women taking the course had a depressive episode in the following year, compared to 25% in the control group<sup>8</sup>.

A meta-analysis of 32 studies in 2014 showed that, in all sorts of groups that are at risk – from expectant and new mothers to individuals who'd experienced a stroke – such preventive interventions reduce the onset of major depressive episodes by 21%, on average<sup>9</sup>. In the same year, my colleagues and I found that 15 of 42 randomized trials reported reductions of 50% or more in the incidence of depression<sup>10</sup>.

Then, this year, the US Preventive Services Task Force reviewed 50 randomized controlled



NACHO DOCE/REUTERS

**Baby yoga in São Paulo, Brazil, which some new mothers find helps them to avoid depression.**

trials testing preventive interventions specifically for perinatal depression. This has shown that, on average, such interventions reduce the incidence of major depressive episodes by 39%. However, one interpersonal approach, called ROSE, reduces the incidence of episodes by 50%, and the Mothers and Babies intervention reduces the incidence of episodes by 53% (ref. 5).

In short, the data suggest that if we implement interventions that seem to be the most effective in clinical trials, we could halve the new cases of major depression.

So why focus on expectant or new mothers? I propose an initial focus on perinatal depression for four reasons. The evidence is strong. The window of risk is clear (during pregnancy and for a year after giving birth). Education and mood-management skills could be wrapped into the prenatal classes or home visits many pregnant women already receive, lowering cost and stigma – as was done in a 2010 study involving more than 2,000 women in the Trent area

of England<sup>11</sup>. Most importantly, interventions could benefit multiple generations. A mother's depression is associated with lower than average birthweight and preterm deliveries, as well as problems in children such as impaired cognitive development<sup>2</sup>. Conversely, the healthy development of babies and children could result in their having healthier, planned pregnancies when they themselves reach childbearing age.

### Making it happen

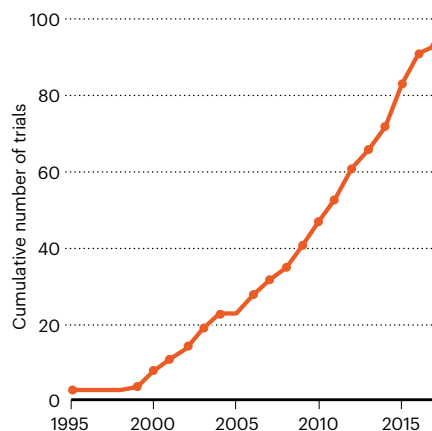
Of course, rolling out evidence-based preventive interventions to millions of women at risk of perinatal depression all over a nation, or the world (many of whom don't have access to prenatal or postnatal care), as well as to other high-risk groups such as adolescents, is a daunting proposition. The number of therapists or health workers available to provide cognitive behavioural courses or interventions based on interpersonal therapy is vastly inadequate, even when it comes to treatment. In 2013, for example, an estimated 43.8 million adults in the United States had experienced a mental illness in the past year, yet only 19.6 million received mental-health services. The World Health Organization estimates that, worldwide, more than 300 million people of all ages experience depression. Most do not receive treatment. A different strategy is required.

In 1998, with support from the Tobacco-Related Disease Research Program in Oakland, California, my colleagues and I began building an online resource to help people to quit smoking. We then conducted a randomized controlled trial in Spanish and English to determine whether people's use of the site could generate quit rates comparable to those obtained from current smoking-cessation aids, such as the nicotine patch. (After six months, the quit rates of people using nicotine patches in the United States are 14–22% (refs 12, 13).)

After registering on the site, people were able to access a guide on how to stop smoking.

### MOUNTING DATA

The number of randomized controlled trials testing preventive interventions for depression has soared.



No source data for years 1996–98 or 2005.

They could submit their 'quit date' and would then receive e-mails advising them on what steps to take as the date drew near. They were given instructions on how to manage their moods, and encouraged to keep diaries as part of the mood-management training. They also became part of an online community that offered support and information.

Our sample consisted of 1,000 smokers from 68 countries, 69% of whom provided follow-up data after one year. (If we didn't get a response to our e-mails or phone calls, we assumed that person had resumed smoking.) In our study, 20% of Spanish speakers and 21% of English speakers quit<sup>14</sup>. In other words, we 'matched the patch'.

Instead of shutting down the website at the end of the grant period, we continued to run the trial with a donation from the Brin Wojcicki Foundation in San Francisco. Over the following 6 years, data from more than 34,000 smokers from 168 countries generated similar results<sup>15,16</sup>. I realized then that our open online interventions were very similar to the now-popular massive open online courses. We had, in fact, carried out a proof-of-concept study of a massive open online intervention, or MOOI<sup>16</sup>.

Various online interventions have already helped to reduce symptoms of depression<sup>17</sup>. Australia's moodgym programme, launched in 2001, is one of the oldest global online interventions for depression. With more than one million registered users, it could well be the most widely used computerized cognitive-behavioural programme in the world<sup>18</sup>. And this month, the UK National Health Service launched an online campaign called Every Mind Matters to help people maintain their mental health.

The interventions I'm calling for would be similar to these, but would need to be built such that their effectiveness could be evaluated on an ongoing basis. Also, effectiveness would need to be made transparent through continually updated 'box scores' on home pages; these could indicate, for instance, that the intervention resulted in a substantial improvement for 20% of 1,000 individuals. People would learn to look for the websites or apps that display effectiveness data, just as they look for services or films that have high ratings.

MOOIs (websites, apps, text-based interventions, and so on) could be provided at no charge to every expectant or new mother in the world – as well as to other groups at risk, such as adolescents, people who have lost a loved one or those who are experiencing physical-health problems. In communities where few people have access to the Internet, health clinics could provide resource rooms where people could access MOOIs. And in remote locations where there are no clinics, local providers could use tablets, laptops or mobile phones to share MOOIs with the people they serve. In fact, the Mothers and Babies course is already being implemented in Tanzania and Kenya.

Some might argue that large-scale

implementation is premature if the risk of developing clinical depression can be cut by only 50%. But such rates of reduction are comparable to those for reducing the risk of influenza through vaccination (40–60%; see [go.nature.com/2wjkr93](http://go.nature.com/2wjkr93)).

Few online and smartphone apps for mental health have been rigorously tested. And some in the field might be concerned that MOOIs could be useless – or worse, harmful. All MOOIs would need to be evidence-based.

Another potential concern is that, once MOOIs become available, insurance companies could refuse to reimburse people for

**“We have the knowledge and the tools to create a world in which fewer people ever experience depression.”**

in-person sessions with counsellors. And there is the worry that MOOIs might exacerbate inequities, with wealthy people receiving cognitive behavioural therapy from therapists and poorer people having access only to online resources. But risks must be weighed against benefits for every intervention.

### Thinking even bigger

So what about the other 50%, whose depression is harder to head off?

As all three NASEM reports<sup>2–4</sup> point out, genetics, other biological factors – such as virus infections or hormone disorders – and people's social and physical environments interact to have a major impact on mental health.

Epigenetics, the study of heritable alterations of the genome structure that are environmentally induced and don't involve changes to the DNA sequence itself, is revealing how life events affect gene expression and the development of mental disorders. Much of the variation in DNA methylation that occurs during the first month of an infant's life, as well as their weight at birth and even some childhood behaviours, have been associated with prenatal environmental factors, such as the mother's smoking habits, mental health and body weight<sup>2</sup>.

The impact of the social environment on development has also been documented in detail. There is growing evidence that nurturing environments (achieved by rewarding good performance at school rather than punishing poor performance, for instance) have a major impact on children's healthy development<sup>19</sup>.

To reach the other 50%, we need to expand beyond individually focused interventions. The 2019 NASEM report<sup>2</sup> recommends that clinical researchers, health-care providers and policy-makers should systematically study and implement more-ambitious interventions that focus on children's social and physical environments.

Soon after arriving at the University of Oregon in Eugene in 1972 to start my doctorate in clinical psychology, I attended a talk at the local community mental-health centre. The speaker chided the professionals in the room, saying, in essence, “We therapists sit in our offices waiting for people to suffer enough to come to see us, or to be brought in by their family or the police because they are being disruptive. We should be going out into the community and sharing what we have learnt so that people can prevent the mental, emotional and behavioural problems that bring them to our offices.”

That evening, I decided to devote much of my professional work to the prevention of mental disorders.

Forty-seven years later, we have the knowledge and the tools to create a world in which fewer people ever experience clinical depression and other mental disorders. Let's start creating it.

### The author

**Ricardo F. Muñoz** is distinguished professor of clinical psychology and founding director of the Institute for International Internet Interventions for Health (i4Health) at Palo Alto University, California; and professor of psychology emeritus in the Department of Psychiatry at the University of California, San Francisco, Zuckerberg San Francisco General Hospital and Trauma Center, California, USA.  
e-mail: [rmunoz@paloaltou.edu](mailto:rmunoz@paloaltou.edu)

1. US Preventive Services Task Force. *J. Am. Med. Assoc.* **321**, 580–587 (2019).
2. National Academies of Sciences, Engineering, and Medicine. *Fostering Healthy Mental, Emotional, and Behavioral Development in Children and Youth: A National Agenda* (Nat'l Acad. Press, 2019).
3. Institute of Medicine. *Reducing Risks for Mental Disorders: Frontiers for Preventive Intervention Research* (Nat'l Acad. Press, 1994).
4. National Research Council & Institute of Medicine. *Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities* (Nat'l Acad. Press, 2009).
5. O'Connor, E., Senger, C. A., Henninger, M. L., Coppola, E. & Gaynes, B. N. *J. Am. Med. Assoc.* **321**, 588–601 (2019).
6. Hasin, D. S. et al. *JAMA Psychiatry* **75**, 336–346 (2018).
7. Mojtabai, R., Olsson, M. & Han, B. *Pediatrics* **138**, e20161878 (2016).
8. Muñoz, R. F. et al. *Cogn. Behav. Pract.* **14**, 70–83 (2007).
9. van Zoonen, K. et al. *Int. J. Epidemiol.* **43**, 318–329 (2014).
10. Muñoz, R. F., Schueller, S. M., Barrera, A. Z., Le, H.-N. & Torres, L. D. in *Handbook of Depression* (eds Gotlib, I. H. & Hammen, C. L.) Ch. 25 (Guilford, 2014).
11. Brugha, T. S., Morrell, C. J., Slade, P. & Walters, S. J. *Psychol. Med.* **41**, 739–748 (2011).
12. Fiore, M. C., Smith, S. S., Jorenby, D. E. & Baker, T. B. *J. Am. Med. Assoc.* **271**, 1940–1947 (1994).
13. Schroeder, S. A. *J. Am. Med. Assoc.* **294**, 482–487 (2005).
14. Muñoz, R. F. et al. *Nicotine Tob. Res.* **11**, 1025–1034 (2009).
15. Leykin, Y., Aguilera, A., Torres, L. D., Pérez-Stable, E. J. & Muñoz, R. F. *J. Med. Internet Res.* **14**, e5 (2012).
16. Muñoz, R. F. et al. *Clin. Psychol. Sci.* **4**, 194–205 (2016).
17. Karyotaki, E. et al. *JAMA Psychiatry* **74**, 351–359 (2017).
18. Twomey, C. & O'Reilly, G. *Aust. N. Z. J. Psychiatry* **51**, 260–269 (2017).
19. Biglan, A. *The Nurture Effect: How the Science of Human Behavior Can Improve Our Lives and Our World* (New Harbinger, 2015).

# Correspondence

## Nobels, gender and ethnicity

As secretary-general of the Royal Swedish Academy of Sciences and secretary of the Nobel Committee for Chemistry, we share your concerns about the shortage of women and of scientists from outside Europe and North America among Nobel laureates (see *Nature* 574, 295; 2019).

However, you are incorrect in saying that we invite only “elite universities and academies” to nominate candidates for the Nobel prize. We make substantial efforts to approach research universities across the world. Each year, we request lists of faculty members from about 200 such universities, selected from some 1,600 institutions on a rotational scheme. We then send individual nomination forms to each professor.

The inequitable distribution of Nobel prizes is a symptom of a bigger problem. Science has been dominated by Western Europe and North America for centuries, and women have had limited scientific opportunities. For example, fewer than 15% of senior authors in *Nature* are women (Y. A. Shen *et al.* *Nature* 555, 165; 2018) and just 2% of your authors are from Africa, South America or western Asia (Nature Index 2019).

Our award process strives to ensure that all scientists get a fair chance, irrespective of geography or gender. As a small contribution, we are launching a programme of Nobel Symposia in Africa. But others must also work to improve the situation – by encouraging women to pursue science careers and by supporting research in low-income countries.

**Göran K. Hansson, Gunnar von Heijne** The Royal Swedish Academy of Sciences, Stockholm, Sweden.  
goran.hansson@kva.se

## India – science and social responsibility

India issued a draft national policy in September for social responsibility in science (see [go.nature.com/32sihv2](https://go.nature.com/32sihv2)). Its aim is to strengthen the country’s knowledge ecosystem, improve communication between science and society, and translate research into social benefits. A central agency and a national digital portal will oversee the policy’s implementation, which is currently being widely discussed in the scientific community.

This policy for scientific social responsibility (SSR) is founded on scientists’ ethical obligation to give back to society in return for the taxpayers’ money that funds their research. It will promote scientific solutions for societal problems such as rural deprivation and the disempowerment of women, and improve scientific and technological support for industry. Funders will be expected to make SSR a condition for awarding grants.

India’s scientific community is crucial for the implementation of the new policy. Researchers will be required to spend a minimum of 10 days every year in public engagement, and to share their knowledge, resources, data sets and equipment to accelerate the advancement of science. Credit for SSR efforts will be given to researchers in their performance evaluations.

Once this policy takes effect, India could lead the way in making science and scientists worldwide more socially responsible.

**Abhay S. D. Rajput** Indian Institute of Tropical Meteorology, Pashan, India.  
abhaysdr@tropmet.res.in

## Chile – right to free will needs definition

Chile could soon become the first country to incorporate ‘neurorights’ into its constitution to prevent the misuse of artificial intelligence and neurotechnology (see [go.nature.com/35qdje5](https://go.nature.com/35qdje5)). These rights relate to the way in which mental processes can be monitored and influenced, principally through brain–computer interfaces.

The country’s Senate has launched a project designed to protect five neurorights proposed by Rafael Yuste, a neurobiologist at Columbia University in New York City (see [go.nature.com/33trrmc](https://go.nature.com/33trrmc) and R. Yuste *et al.* *Nature* 551, 159–163; 2017). These include the right to ‘free will’. However, the precise meaning of free will should be carefully debated before it is incorporated as a national or international right.

Free will is a multidimensional concept that poses several unsolved philosophical problems. Most prominent is whether free will is compatible with determinism. Cultural diversity can also influence its interpretation (see N. Chernyak *et al.* *Dev. Psychol.* 55, 866–876; 2019). Furthermore, different types of free choice and action have already been included in the Universal Declaration of Human Rights (see articles 16.2, 18 and 21.3; [go.nature.com/33t9bhn](https://go.nature.com/33t9bhn)).

We should therefore work towards developing a consensual, minimal definition of free will. Although such a definition might not close the philosophical debate, it could be ethically operational in that it would help to pre-empt misinterpretations based on legislative loopholes.

**José M. Muñoz** European University of Valencia, Spain.  
josemanuel.munoz@universidadeuropea.es

## Test reproducibility of old computer code

We question whether analytical tools such as Common Workflow Language, which aim to make computational methods “reproducible and shareable”, can stand the test of time (see *Nature* 573, 149–150; 2019). The long-term validity of computational results will not be testable if the original code cannot be run many years later.

Considering the rapidity of transformations in operating systems and programming languages, it is hard to predict the lifetime reproducibility of a particular code. We have therefore organized the Ten Years Reproducibility Challenge (see [go.nature.com/2bwcukq](https://go.nature.com/2bwcukq)). Researchers are invited to test code reproducibility by trying to rerun a code created for a scientific paper they published more than ten years ago. The codes can address any scientific domain (statistical analysis, numerical simulation or data processing, for example) and be written in any language.

The challenge closes in April 2020. Our hope is that the results will offer insights into long-term causes of non-reproducibility.

**Konrad Hinsen** Centre for Molecular Biophysics, CNRS, Orléans, France.

**Nicolas Rougier** Inria Bordeaux, Talence, France.  
nicolas.rougier@inria.fr

## HOW TO SUBMIT

Correspondence may be submitted to [correspondence@nature.com](mailto:correspondence@nature.com) after consulting the author guidelines and section policies at [go.nature.com/cmchno](https://go.nature.com/cmchno).



# News & views

## Cell biology

# Senescent cells feed on their neighbours

Michael Overholtzer

Chemotherapy-treated cancer cells that enter a non-dividing state called senescence can nevertheless boost cancer growth. The finding that these cells eat neighbouring cells reveals a mechanism that enables senescent cells to persist.

Multicellular life requires individual cells to cooperate in a way that benefits the organism. Cells that are uncooperative because they are damaged or dysfunctional, and that pose a threat, are either eliminated by cell death or undergo a usually irreversible growth arrest called senescence<sup>1</sup>. Senescent cells typically never divide (although there are some rare examples of cells exiting senescence and resuming division), but they can persist in tissues and contribute to ageing and cancer progression<sup>2,3</sup>. Writing in the *Journal of Cell*

*Biology*, Tonnessen-Murray *et al.*<sup>4</sup> reveal a deadly activity that underlies the persistence of senescent cells – they can eat their neighbours alive.

Cellular entry into senescence benefits an organism because it inhibits cancer development by preventing the division of cells that have accumulated extensive DNA damage or that express cancer-promoting genes called oncogenes<sup>2,5</sup>. Senescent cells are metabolically active<sup>6</sup>, and this is characterized by their secretion of pro-inflammatory

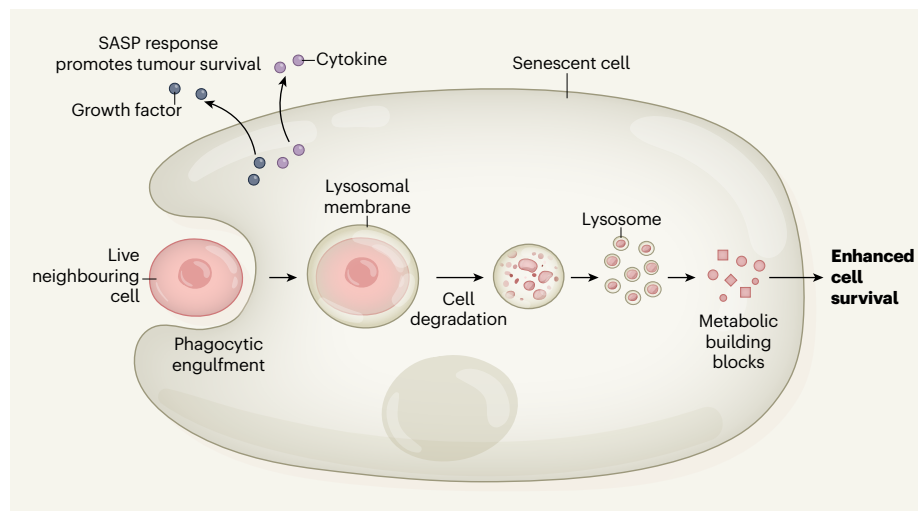
molecules as part of a phenomenon termed the senescence-associated secretory phenotype (SASP) response<sup>2</sup>. Senescent cells can promote cancer progression and resistance to anticancer therapy in some contexts, as a result of the secretion, through SASP, of growth factors and immune-signalling molecules called cytokines<sup>2</sup>.

Chemotherapy that damages the DNA of cancer cells can result in their death or their entry into senescence. Tonnessen-Murray and colleagues investigated the effects of chemotherapy-driven senescence in breast cancer cells in mice treated with the chemotherapeutic drug doxorubicin. Under the microscope, they saw senescent cells eating and digesting entire neighbouring cells (Fig. 1). This striking observation was made in breast tumours formed of mixtures of transplanted cancer cells, which were engineered to express red or green fluorescent proteins. It can be difficult to observe a cell being internalized by another cell (a process termed engulfment) in cancer tissues. By growing tumours with mixtures of fluorescently labelled cells, the authors could clearly identify red- or green-labelled cells being taken up into neighbouring cells labelled by the other colour.

Engulfment also occurred at high rates for mouse and human breast cancer cells grown *in vitro* and treated with doxorubicin or another chemotherapeutic drug, paclitaxel. Ingestion peaked at 4–6 days after drug treatment, a time that correlated with the induction of senescence. The cells that were engulfed by the senescent cells were neighbouring senescent or non-senescent cancer cells. They showed no sign of being dead, and engulfment occurred even in the presence of a cell-death inhibitor molecule. This led the authors to conclude that the ingested cells were being eaten alive.

Ingested cells are broken down in a digestive organelle called the lysosome. Crucially, senescent cells that ate their neighbours survived longer *in vitro* than those that did not. This finding suggests that metabolic building blocks retrieved from the lysosomal digestion of neighbouring cells were being used by senescent cells to promote their survival.

This surprising finding that cell death supports the survival of senescent cells highlights the complexity of cell-death regulation in multicellular animals. Numerous mechanisms of cell death occur in animal tissues. These include forms of cell suicide, such as apoptosis, which leads to the fragmentation of individual cells, and regulated forms of necrotic cell death that induce cell rupture<sup>7</sup>. Some cell deaths are also carried out as ‘murders’<sup>7,8</sup>. These typically



**Figure 1 | Cellular cannibalism.** Chemotherapy drugs can cause cancer cells to enter a state of senescence, which is usually associated with an irreversible halt to cell division. However, these cells can promote tumour survival by secreting growth factors and signalling molecules called cytokines in a process termed the senescence-associated secretory phenotype (SASP) response<sup>2</sup>. Tonnessen-Murray *et al.*<sup>4</sup> report studies of breast cancer in mice which reveal that this type of senescent cell takes up (engulfs) and digests neighbouring living cells. The cells are engulfed by a process that has molecular characteristics of phagocytosis, an engulfment process that immune cells use. Once ingested, the cells are enveloped in membrane from an organelle called the lysosome and digested. This might account for a portion of the numerous lysosomes that are a hallmark of senescent cells. This degradation provides metabolic building blocks for the cell. Senescent cells that have ingested their neighbours survive for longer than senescent cells that have not.

involve the presence of engulfing cells, and occur by at least two distinct mechanisms<sup>9</sup>.

One is a form of cell death called entosis, in which living cells that are destined to die invade a neighbouring cell and become engulfed<sup>10</sup>. Another mechanism is cellular cannibalism, in which living cells that will be ingested are targeted by a type of engulfment that resembles phagocytosis – the process typically used by immune-system cells such as macrophages to ingest and destroy dying cells<sup>9</sup>. Such cellular murders can support the survival of particular cells in a population that benefit from the metabolic banquet derived from ingesting and degrading whole cells<sup>11,12</sup>.

The authors examined the mechanism of senescence-associated engulfment and found that, although entosis could occur in the type of tumour cell studied, the engulfment of senescent cells did not involve the proteins required for entosis<sup>10</sup>. The authors analysed the gene-expression profile of cancer cells treated with chemotherapy drugs (most of these cells were senescent), and found that genes characteristic of phagocytosis were expressed. This gene expression peaked within a timeframe that correlated with the cellular engulfment. Senescent cells were also observed to engulf dead cells added *in vitro*, providing further evidence for the authors' model that senescent cells engulf cells by phagocytosis.

Cell cannibalism in cancers has been reported previously<sup>9,12</sup>. However, Tonnessen-Murray *et al.* specifically identify an association between cannibalism and senescence, and show that this phenomenon might make a substantial contribution to the persistence of senescent cells in cancer tissues. The authors observed that cannibalism by senescent breast cancer cells occurs irrespective of whether or not the cell has functional p53, a notable tumour-suppressor protein that can control entry into senescence<sup>13</sup>. The authors tested chemotherapy-induced senescent cells of other types of cancer, including lung cancer and a bone cancer called osteosarcoma, and found that these cells also cannibalize neighbouring cells. Together, these findings suggest that cell cannibalism might be an activity that is broadly associated with the induction of senescence, rather than being linked to particular types of cancer or to the status of proteins such as p53. It will be important to investigate whether cannibalism is linked to senescence in other contexts, for example during tissue development when senescence can occur<sup>14,15</sup>, or in aged tissues that accumulate senescent cells<sup>3</sup>.

Entosis in cancer-cell populations can promote competition between individual cells in which 'winner' cells ingest and kill neighbouring 'loser' cells, removing them from the population<sup>16</sup>. Whether cells behave as winners or losers depends on certain cellular characteristics, for example differences in the tension of the internal cellular framework called the

cytoskeleton<sup>16</sup>. It would be interesting to investigate whether senescent cells choose particular target cells to cannibalize in a competitive fashion. In cancers, complex mixtures of cells coexist in the tumour microenvironment, and this cellular composition changes over time or in response to anticancer therapy. The authors propose that cell cannibalism might affect cancer progression by supporting the SASP response. However, it is worth considering whether it might also contribute directly to cancer progression by removing particular cells from the tumour microenvironment. And if normal cells are found to be removed by senescent cells in aged tissues, this depletion might contribute directly to tissue degeneration.

**Michael Overholtzer** is in the Cell Biology Program, Memorial Sloan Kettering Cancer

Center, New York, New York 10065, USA.  
e-mail: overhom1@mskcc.org

- Herranz, N. & Gil, J. *J. Clin. Invest.* **128**, 1238–1246 (2018).
- Faget, D. V., Ren, Q. & Stewart, S. A. *Nature Rev. Cancer* **19**, 439–453 (2019).
- McHugh, D. & Gil, J. *J. Cell Biol.* **217**, 65–77 (2018).
- Tonnessen-Murray, C. A. *et al. J. Cell Biol.* <https://doi.org/10.1083/jcb.201904051> (2019).
- Hayflick, L. & Moorhead, P. S. *Exp. Cell Res.* **25**, 585–621 (1961).
- Quijano, C. *et al. Cell Cycle* **11**, 1383–1392 (2012).
- Galluzzi, L. *et al. Cell Death Differ.* **25**, 486–541 (2018).
- Martins, I. *et al. Biomed. J.* **40**, 133–140 (2017).
- Fais, S. & Overholtzer, M. *Nature Rev. Cancer* **18**, 758–766 (2018).
- Overholtzer, M. *et al. Cell* **131**, 966–979 (2007).
- Hamann, J. C. *et al. Cell Rep.* **20**, 201–210 (2017).
- Lugini, L. *et al. Cancer Res.* **66**, 3629–3638 (2006).
- Shay, J. W., Pereira-Smith, O. M. & Wright, W. E. *Exp. Cell Res.* **196**, 33–39 (1991).
- Munoz-Espin, D. *et al. Cell* **155**, 1104–1118 (2013).
- Storer, M. *et al. Cell* **155**, 1119–1130 (2013).
- Sun, Q. *et al. Cell Res.* **24**, 1299–1310 (2014).

## Palaeoclimate

# Fresh evidence in the glacial-cycle debate

**Eric W. Wolff**

An analysis of air up to 2 million years old, trapped in Antarctic ice, shows that a major shift in the periodicity of glacial cycles was probably not caused by a long-term decline in atmospheric levels of carbon dioxide. **See p.663**

During the past 2.6 million years, Earth's climate has alternated between warm periods known as interglacials, when conditions were similar to those of today, and cold glacials, when ice sheets spread across North America and northern Europe. Before about 1 million years ago, the warm periods recurred every 40,000 years, but after that, the return period lengthened to an average of about 100,000 years. It has often been suggested that a decline in the atmospheric concentration of carbon dioxide was responsible for this fundamental change. On page 663, Yan *et al.*<sup>1</sup> report the first direct measurements of atmospheric CO<sub>2</sub> concentrations from more than 1 million years ago. Their data show that, although CO<sub>2</sub> levels during glacials stayed well above the lows that occurred during the deep glacials of the past 800,000 years, the maximum CO<sub>2</sub> concentrations during interglacials did not decline. The explanation for the change must therefore lie elsewhere.

Understanding what caused the shift in periodicity, known as the mid-Pleistocene transition (MPT), is one of the great challenges of palaeoclimate science. The 40,000-year periodicity that dominated until about 1 million years ago is easily explained, because the tilt

of Earth's spin axis relative to its orbit around the Sun varies between 22.1° and 24.5° with the same period. In other words, before the MPT, low tilts led to cooler summers that promoted the growth and preservation of ice sheets.

But after the MPT, glacial cycles lasted for two to three tilt cycles. Because the pattern of variation in Earth's orbit and tilt remained unchanged, this implies that the energy needed to lose ice sheets<sup>2</sup> had increased. One prominent explanation<sup>3</sup> is that atmospheric levels of CO<sub>2</sub> were declining, and eventually crossed a threshold value below which the net cooling effect of the decline allowed ice sheets to persist and grow larger.

Ancient air trapped in Antarctic ice can be extracted from cores drilled from the ice sheet, allowing the CO<sub>2</sub> concentration to be measured directly, but the ice-core record extends to only 800,000 years ago<sup>4</sup>. Estimates of CO<sub>2</sub> concentrations from earlier periods have been made by measuring the ratio of boron isotopes in shells found in ancient marine sediments<sup>5,6</sup>. This proxy measurement depends on a chemical equilibrium controlled by ocean acidity, which, in turn, is closely related to the atmospheric CO<sub>2</sub> concentration.



But the estimates of CO<sub>2</sub> levels inferred from such measurements are necessarily imprecise and must be verified using more-precise, direct measurements. Scientists have therefore formulated plans<sup>7</sup> to find and retrieve deep ice cores that reach back to before the MPT (see [go.nature.com/33mw4yk](https://go.nature.com/33mw4yk)). One project has recently been funded by the European Union, and hopes to retrieve million-year-old ice in 2024.

Yan *et al.* tried another approach to finding similarly old ice, but nearer the surface of Antarctica. In regions known as blue-ice areas, the combination of ice flow against a mountain barrier and surface ice loss by wind scouring and sublimation (transformation of ice directly into water vapour) leads to upwelling of old ice towards the surface. The authors therefore studied two cores, 147 and 191 metres deep, that were drilled to bedrock in the blue-ice region near the Allan Hills in Antarctica (Fig. 1).

The researchers improved and applied a relatively new method<sup>8</sup> to date this old ice. The concentration of argon-40 in Earth's atmosphere is slowly increasing with time as it is produced from the radioactive decay of potassium-40. By measuring the ratios of argon isotopes in air extracted from cores, the age of ice can be determined. The authors also measured the ratios of deuterium (a heavy isotope of hydrogen) to hydrogen in the ice, which can be used as a proxy of temperature at the time the ice was deposited.

Yan and colleagues concluded that ice in the lowest 30 m of each core is up to 2.7 million years old. However, the uncertainty of 100,000 years in this dating precludes their samples from being matched to particular parts of Earth's tilt cycle. Moreover, the authors found abrupt age discontinuities with depth in the cores, which suggests that the layers of ice within them have been disturbed. The authors therefore treated the measured concentrations of deuterium and CO<sub>2</sub> as snapshots of climate and atmospheric composition that corresponded to an approximate age of the ice, rather than as an ordered time series. On the basis of the deuterium values, they make a plausible case that the observed range of measured CO<sub>2</sub> values represents most of the actual glacial–interglacial range.

Unfortunately, in the oldest ice samples, there was evidence that the CO<sub>2</sub> concentration had been artificially enhanced by gas produced from the breakdown of organic material at the base of the ice sheet. A few samples from about 2 million years ago were potentially not affected by this issue, but were insufficient in number to allow any conclusions to be drawn about the range of CO<sub>2</sub> levels at that time.

However, Yan *et al.* obtained samples from about 1 million and 1.5 million years ago that they consider to be undisturbed by the artificial addition of CO<sub>2</sub>. In both periods, the maximum CO<sub>2</sub> concentrations are



**Figure 1 | Blue ice near the Allan Hills region of Antarctica.** The environmental conditions in this area draw ancient ice to the surface. Yan *et al.*<sup>1</sup> have analysed air trapped in an ice core drilled from this region to obtain the first direct measurements of atmospheric carbon dioxide levels from more than 1 million years ago.

similar to those of interglacials from the past 500,000 years, peaking at 279 parts per million (p.p.m.). But the minimum value of 214 p.p.m. is much higher than the lows of around 180 p.p.m. that occurred during recent glacial maxima (the periods that corresponded to the maximum extent of ice).

The authors conclude that the relationship between CO<sub>2</sub> levels and Antarctic temperature was similar before and after the MPT. The fact that the pre-MPT ice does not contain very low

**“These data force us to look elsewhere for the cause of the longer glacial cycles.”**

ratios of deuterium to hydrogen that would be characteristic of extremely cold Antarctic temperatures, nor low CO<sub>2</sub> levels characteristic of recent glacial maxima, is probably just a consequence of the shorter period of the glacial cycles. Such low values are generally not found in the first 40,000 years of post-MPT glacial cycles either.

Although Yan and colleagues' data points cannot be placed within a tilt cycle, it seems likely that the CO<sub>2</sub> concentrations are not very different at the crucial points in cycles when the ice sheet is either lost (before the MPT) or continues growing (after the MPT). This forces

us to look elsewhere for the cause of the longer cycles, perhaps refocusing efforts on understanding whether changes to the nature of the ice-sheet bed caused by glacial erosion<sup>9</sup> altered the characteristics of the ice sheets and their vulnerability to melting.

Yan and colleagues' data add much-needed precision to the previously reported estimates of CO<sub>2</sub> levels made using data from marine sediments<sup>5,6</sup>. However, their tantalizing snapshots of the pre-MPT world emphasize the need for a complete, undisturbed time series of greenhouse-gas concentrations that can be put into context with the climate cycles at that time. Let us hope that the planned new ice cores will provide that.

**Eric W. Wolff** is in the Department of Earth Sciences, University of Cambridge, Cambridge CB2 3EQ, UK.  
e-mail: [ew428@cam.ac.uk](mailto:ew428@cam.ac.uk)

1. Yan, Y. *et al.* *Nature* **574**, 663–666 (2019).
2. Tzedakis, P. C., Crucifix, M., Mitsui, T. & Wolff, E. W. *Nature* **542**, 427–432 (2017).
3. Berger, A., Li, X. S. & Loutre, M. F. *Quat. Sci. Rev.* **18**, 1–11 (1999).
4. Bereiter, B. *et al.* *Geophys. Res. Lett.* **42**, 542–549 (2015).
5. Chalk, T. B. *et al.* *Proc. Natl Acad. Sci. USA* **114**, 13114–13119 (2017).
6. Dyez, K. A., Hönisch, B. & Schmidt, G. A. *Paleoceanogr. Paleoclimatol.* **33**, 1270–1291 (2018).
7. Fischer, H. *et al.* *Clim. Past* **9**, 2489–2505 (2013).
8. Bender, M. L., Barnett, B., Dreyfus, G., Jouzel, J. & Porcelli, D. *Proc. Natl Acad. Sci. USA* **105**, 8232–8237 (2008).
9. Clark, P. U. *et al.* *Quat. Sci. Rev.* **25**, 3150–3184 (2006).



## Microbiology

# Bacterial twist to an antiviral defence

Karen L. Maxwell

The discovery of an antiviral defence system in bacteria that shares some components with a key antiviral defence pathway in animals provides insight into how this important response might have evolved. **See p.691**

Humans face a daily threat of infection by harmful viruses. To repel them, our immune system mounts an immediate response following invasion that depends on its ability to recognize general characteristics indicating that viruses are foreign entities. This type of reaction, generated by an ancient branch of the immune system known as innate immunity, occurs in all plants and animals. Many genes involved in innate immune responses are evolutionarily conserved and encode proteins that are used for defence purposes in different species<sup>1–3</sup>. Cohen *et al.*<sup>4</sup> report on page 691 that some bacterial species fight viral infections by using an innate immune mechanism that is related to one of the central components of innate immunity in animals called the cGAS–STING pathway. Their findings reveal that this crucial antiviral defence system in animals might have its evolutionary roots in bacteria.

There has been a rise in evidence indicating that the defence systems mediating innate immunity in animals have counterparts in bacteria. For example, protein components called TIR domains, which are present in defence proteins in mammals and plants, can recognize molecular hallmarks of disease-causing agents known as PAMPs, and then trigger an immune response. TIR domains are evolutionarily conserved in bacteria, protecting them from viruses called phages<sup>5</sup>. Another such example is the antiviral machinery that targets RNA and depends on proteins called Argonautes, found in plants and animals. This system also has a role in defence responses in bacteria and the single-celled organisms known as archaea<sup>6,7</sup>.

The evolutionary conservation of these innate immune mechanisms in bacteria and mammals suggests that such pathways might have first arisen in bacteria as protection

against phages, and have since evolved into different, but related, defences across the tree of life. One key open question is how many of the innate immune defences found in animals might have evolved from ancient bacterial systems.

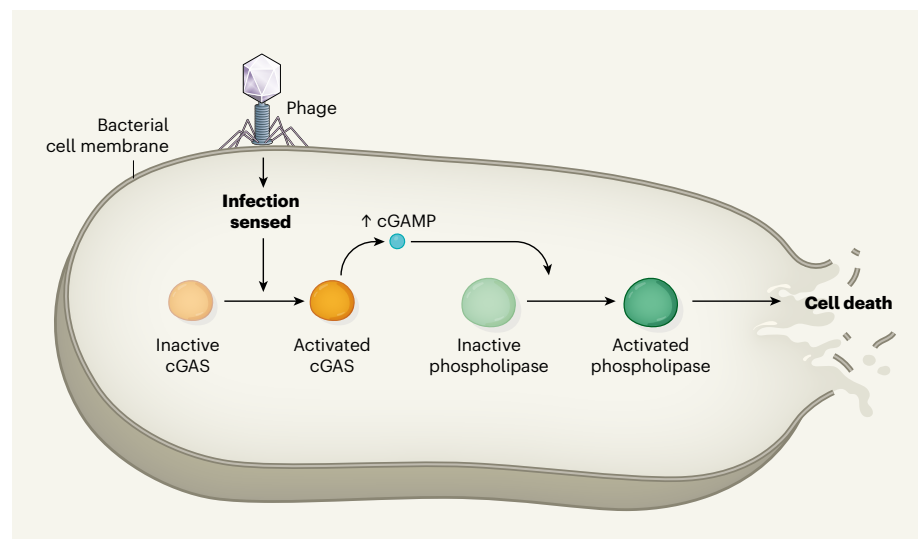
When the cGAS–STING pathway<sup>8,9</sup> in animals detects invading viruses in a cell, it activates a response that either mediates antiviral defences or triggers cell death<sup>10</sup>. The cGAS enzyme functions in this defence by sensing and binding to double-stranded viral DNA, and then inducing<sup>11</sup> the production of a type of signalling molecule called cGAMP, which is termed a cyclic dinucleotide. The binding of cGAMP to the STING protein sets off a signalling cascade that unleashes an antiviral response.

Cohen and colleagues analysed regions of bacterial genomes in which defence genes are clustered, and noticed that the gene encoding cGAS was often located near genes whose products involved in other antiphage defence systems, such as CRISPR–Cas. The authors therefore wondered whether cGAS might have a role in antiphage defences.

To test this idea, Cohen *et al.* engineered bacteria lacking a cGAS system to express genes encoding such systems. The authors tested two representative cGAS systems (comprising the gene encoding cGAS and three adjacent genes) from the bacterial species *Vibrio cholerae* and *Escherichia coli*. Both cGAS systems conferred a resistance to infection by diverse phages. When the authors disrupted the DNA sequence of the cGAS-system genes, resistance to phage infection was completely lost – confirming that bacteria use this cGAS signalling pathway for antiviral defence, much as do eukaryotes (multicellular organisms that have a nucleus in their cells). The authors called this antiphage defence system cyclic-oligonucleotide-based anti-phage signalling system (CBASS). Genes encoding cGAS proteins are present in approximately 10% of all sequenced bacterial genomes<sup>12</sup>, suggesting that CBASS systems have a widespread role in antiphage defences.

The pathway downstream of cGAMP production in bacteria differs from that in animals. The authors report that, in bacteria, cGAMP production activated a phospholipase enzyme in some CBASS systems (Fig. 1). This activated phospholipase then degraded phospholipid molecules in the bacterial cell membrane, killing the infected bacterium. Such cellular ‘suicide’ could protect a bacterial cell population because the destruction of infected cells through this process prevents the phage from spreading to neighbouring bacteria.

In some bacterial species, the CBASS defence systems identified by Cohen and colleagues lacked a phospholipase component. These systems instead encoded proteins that might trigger cell suicide through alternative



**Figure 1 | A bacterial defence against viral infection.** Cohen *et al.*<sup>4</sup> report a defence system that is used by bacteria to fight infection by bacterium-infecting viruses called phages. After phage infection is sensed (through an unknown mechanism), the enzyme cGAS is activated, producing the molecule cGAMP. Such changes in cGAS activity and cGAMP levels in response to viral infection also occur in a range of multicellular organisms, including humans. In the bacterial system, the rise in cGAMP can lead to activation of a phospholipase enzyme that degrades phospholipid molecules in the bacterial cell membrane. This process kills the bacterial cell and can stop viral infection from spreading to neighbouring cells.

mechanisms, such as degrading the bacterial genome or creating a hole in the cell membrane through the action of a pore-forming protein. But whether these systems kill cells in such ways remains to be tested. In some cases, the CBASS systems encoded a protein in which a TIR domain was fused to a STING domain similar to that in eukaryotes. The evolutionary conservation of these domains in an antiviral defence system in bacteria suggests that they might represent the ancient evolutionary origin of the eukaryotic cGAS–STING defence system.

Although some CBASS systems had only cGAS genes and components required for bacterial cell death, others had genes whose products were associated with ubiquitination, a protein-modification pathway in eukaryotic cells. In this process, a protein called ubiquitin is attached to a target by an enzyme-mediated reaction. CBASS systems included proteins that have several components associated with eukaryotic ubiquitination: E1 and E2 domains, typically found in enzymes that mediate ubiquitin activation and transfer, respectively, and JAB domains, which are found in proteins that remove ubiquitin from targets. Ubiquitination fine-tunes the length and intensity of innate immune responses in animals<sup>13</sup>. This provides yet another link connecting bacterial and animal antiviral responses. The ubiquitination components of the *E. coli* CBASS system were required for defence against some but not all phages, suggesting that these proteins might allow systems to recognize specific phage proteins or features, rather than being a more general property of phages – thereby refining the activity of these systems.

Antiphage defence systems in bacteria can be a target of phage-encoded inhibitor proteins. For example, phage proteins can block CRISPR–Cas defences<sup>14</sup>. It is highly probable that some phages have evolved ways to inhibit CBASS systems. Different CBASS systems encode a diverse set of cyclic-oligonucleotide signalling molecules and components, suggesting that cell suicide occurs through a number of mechanisms. The diversity of these CBASS-system components is probably driven by the need to evade a phage counter-attack if, for example, a phage-encoded protein could inactivate a particular cyclic-oligonucleotide signalling molecule. The selective pressure from antiphage systems that phages encounter would inevitably lead to the evolution of countermeasures in these viruses. An exciting area for future research will be to search for such phage inhibitors of CBASS systems.

One key aspect of cGAS function in bacterial defence that remains unknown is which signal the immune system detects to recognize that a viral infection is occurring. In eukaryotes, any viral double-stranded DNA in the cytoplasm can be recognized as a foreign entity because eukaryotic DNA is usually confined to the nucleus and absent from the

cytoplasm. To distinguish cytoplasmic viral DNA from bacterial DNA, a bacterium lacking a nucleus would presumably require a sensor with a nuanced capacity to identify foreign DNA. One possibility is that CBASS systems recognize phage DNA specifically in the linear, relaxed state that occurs immediately after it has entered the bacterial cell. Perhaps the proteins that have E1, E2 and JAB domains in CBASS systems provide further refinement to aid the success of this aspect of phage recognition.

Cohen and colleagues' study is particularly remarkable for highlighting the striking parallels between innate immunity in eukaryotes and bacteria. The number of known bacterial antiphage systems is growing rapidly<sup>5,15,16</sup>, and it is probable that many more such exciting connections remain to be uncovered.

**Karen L. Maxwell** is in the Department of Biochemistry, University of Toronto, Toronto,

Ontario M5G 1M1, Canada.  
e-mail: karen.maxwell@utoronto.ca

1. Travis, J. *Science* **324**, 580–582 (2009).
2. Jones, J. D., Vance, R. E. & Dangl, J. L. *Science* **354**, aaf6395 (2016).
3. Hoffmann, J. & Akira, S. *Curr. Opin. Immunol.* **25**, 1–3 (2013).
4. Cohen, D. et al. *Nature* **574**, 691–695 (2019).
5. Doron, S. et al. *Science* **359**, eaar4120 (2018).
6. Swarts, D. C. et al. *Nature* **507**, 258–261 (2014).
7. Olovnikov, I., Chan, K., Sachidanandam, R., Newman, D. K. & Aravin, A. A. *Mol. Cell* **51**, 594–605 (2013).
8. Kranzusch, P. J. et al. *Mol. Cell* **59**, 891–903 (2015).
9. Margolis, S. R., Wilson, S. C. & Vance, R. E. *Trends Immunol.* **38**, 733–743 (2017).
10. Ma, Z. & Damanian, B. *Cell Host Microbe* **19**, 150–158 (2016).
11. Sun, L., Wu, J., Du, F., Chen, X. & Chen, Z. J. *Science* **339**, 786–791 (2013).
12. Whiteley, A. T. et al. *Nature* **567**, 194–199 (2019).
13. Liu, J., Qian, C. & Cao, X. *Immunity* **45**, 15–30 (2016).
14. Stanley, S. Y. & Maxwell, K. L. *Annu. Rev. Genet.* **52**, 445–464 (2018).
15. Goldfarb, T. et al. *EMBO J.* **34**, 169–183 (2015).
16. Ofir, G. et al. *Nature Microbiol.* **3**, 90–98 (2018).

This article was published online on 8 October 2019.

## Cancer

# Teamwork by T cells boosts immunotherapy

Jonathan L. Linehan & Lélia Delamarre

Immunotherapy treatment harnesses CD8 T cells of the immune system to kill tumour cells. The finding that CD4 helper T cells contribute to the success of this treatment in mice might offer a way to improve clinical outcomes. **See p.696**

Immune cells called CD8 (or cytotoxic) T cells can target and kill cancer cells, and immunotherapies that boost this process are in clinical use. However, for reasons that are not fully clear, it is hard to predict whether a person will respond to this treatment. On page 696, Alspach *et al.*<sup>1</sup> report mouse studies revealing that another type of immune cell, called a CD4 cell (also known as a helper T cell), has a crucial role in aiding CD8 T cells to target tumours after immunotherapy.

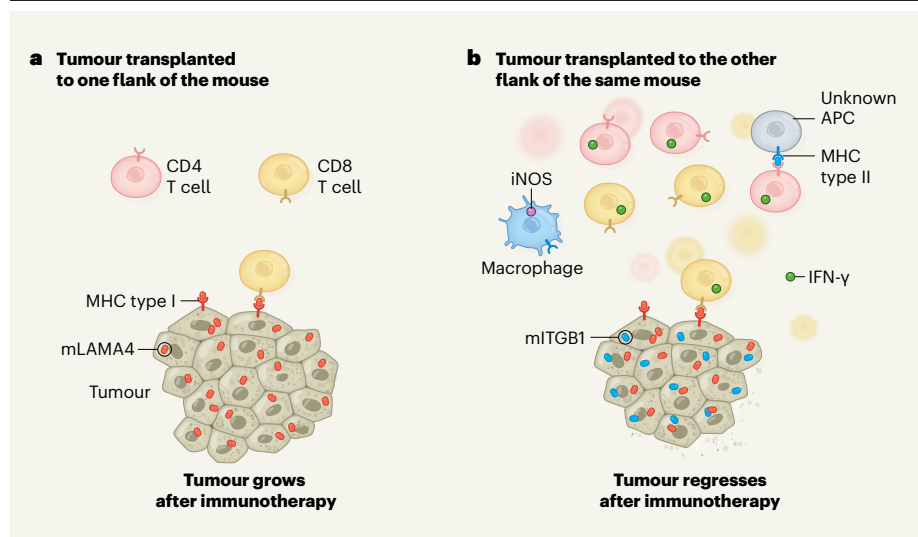
Mutations in tumour cells can give rise to abnormal proteins, fragments of which – termed neoantigens – are displayed on the surface of cells bound to major histocompatibility complex (MHC) molecules. If a neoantigen is recognized by a CD8 T cell, this cell can target and kill any tumour cells that express the neoantigen. However, this cytotoxic response can be blocked, for example by an immunosuppressive environment surrounding a tumour. Immunotherapy treatments called immune-checkpoint blockade or immune-checkpoint therapy can counteract such problems to enable CD8 T cells to unleash an effective immune response against the tumour.

Much immunotherapy research focuses on CD8 T cells. However, there is emerging evidence that CD4 T cells might have a key role in tumour-targeting immune responses<sup>2,3</sup>.

Alspach and colleagues sought to identify the minimal immune-stimulating neoantigen requirement to drive an effective immune response in mice that were given an immunotherapy treatment. The authors studied mice that had a type of tumour to which the immune system does not normally respond, and they engineered such tumours to express neoantigens. The neoantigen termed mLAMA4 is recognized by CD8 T cells<sup>4</sup>, and the neoantigen termed mITGB1, recognized by CD4 T cells, was identified by the authors using a computational prediction method. In the absence of immunotherapy, the expression of these two neoantigens, either alone or together in a tumour, was insufficient to trigger an effective immune response against the tumour. However, if both neoantigens were expressed in animals receiving immunotherapy, the tumour regressed.

To check whether this response was simply





**Figure 1 | T-cell collaboration drives an effective anti-tumour response to immunotherapy.** Alspach *et al.*<sup>1</sup> studied how mice responded to immunotherapy that boosts tumour destruction by CD8 T cells. The mice received transplanted tumours in each flank that expressed abnormal protein fragments called neoantigens. The animals were then treated with immunotherapy. The neoantigen mLAMA4 is presented by type I major histocompatibility complex (MHC) molecules and recognized by immune cells called CD8 T cells, and the neoantigen mITGB1 is presented by type II MHC molecules and recognized by other immune cells called CD4 T cells. **a**, A tumour that expressed mLAMA4 attracted few immune cells and grew. **b**, By contrast, a tumour on the animal's opposite flank that expressed mLAMA4 and mITGB1 regressed, indicating the importance of activating both CD8 and CD4 T cells at the tumour site to generate a successful response to immunotherapy. The robust immune response generated included the accumulation of CD4 and CD8 T cells (which produce the signalling protein IFN- $\gamma$ ) and macrophage cells (which expressed the protein iNOS). The tumour cells lacked type II MHC molecules; therefore, an as-yet-unidentified antigen-presenting cell (APC) probably presents mITGB1 to CD4 T cells.

dependent on neoantigen quantity, rather than the need for neoantigen recognition by both types of immune cell, the authors engineered mouse tumours to express two different neoantigens that are recognized by CD8 T cells, but not by CD4 cells. These tumours did not respond to immunotherapy, demonstrating that a successful immune response depends on the presence of neoantigens that trigger responses from both CD4 and CD8 T cells.

The authors' analysis reveals that the CD4 T cells that responded to mITGB1 had the hallmarks of a type of CD4 T cell called a T helper type 1 cell, which can increase the number and cell-killing activity of CD8 T cells<sup>2</sup>. The authors confirmed that, if tumours expressed both mLAMA4 and mITGB1, this indeed caused an increase in the number and cytotoxic activity of CD8 T cells, compared with the case for animals with tumours that expressed only mLAMA4. Alspach and colleagues also showed that, if animals were first vaccinated with dying tumour cells and were then implanted with a growing tumour that expressed mLAMA4 and mITGB1, the transplanted tumours were most efficiently rejected if the vaccine contained tumour cells that expressed both mLAMA4 and mITGB1 in the same cell.

To determine whether CD4 T cells have a role beyond just enhancing the priming of CD8 T cells, as occurs during vaccination, the authors investigated whether mITGB1 is required at the

tumour site for an active immune response. They implanted mice with a tumour that expressed both mLAMA4 and mITGB1 on one flank and with a tumour that expressed only mLAMA4 on the opposite flank, and treated the mice with immunotherapy (Fig. 1). As expected, the tumour that expressed mLAMA4 and mITGB1 was targeted by the immune system and regressed, but the tumour that expressed only mLAMA4 continued to grow slowly. In comparison with the mLAMA4- and mITGB1-expressing tumour, the growing tumour was infiltrated by fewer CD4 T cells, and by fewer CD8 T cells that could recognize mLAMA4.

These results highlight the need for a tumour to express neoantigens that are recognized by both CD4 and CD8 T cells to generate a productive response to immunotherapy. Together, these data demonstrate that CD4 T cells not only aid the priming of CD8 T cells, but also collaborate with CD8 T cells at the tumour site to maintain an effective anti-tumour response during immunotherapy. The mechanism enabling this collaboration remains to be determined.

The authors suggest that interferon- $\gamma$  (IFN- $\gamma$ ), a type of immune-signalling protein called a cytokine, might be one necessary component enabling this collaboration. IFN- $\gamma$  is produced by CD8 T cells and CD4 helper T cells, and can help to tackle an immunosuppressive tumour environment<sup>3</sup>.

Tumours that are responsive to immunotherapy are associated with the presence of activated immune cells called macrophages that express the protein iNOS (ref. 6). Alspach and colleagues observed that tumours that expressed mLAMA4 and mITGB1 had an impressive 83-fold increase in the presence of iNOS-expressing macrophages in comparison with tumours that expressed only mITGB1. CD4 helper T cells alone are not sufficient to drive this iNOS expression in macrophages, and CD8 T cells are also required, suggesting an interplay between these three types of cell. Consistent with these findings, previous work<sup>7</sup> indicates that macrophage activation triggered by IFN- $\gamma$  from CD4 T cells leads to the inhibition of tumour growth.

The tumour cells studied by the authors express type I MHC molecules that present neoantigens to CD8 T cells, but they do not express type II MHC molecules that present neoantigens to CD4 T cells. The identification of the immune cells that present neoantigens, such as mITGB1, to CD4 T cells in this system should be a topic for future research. Antigen-presenting immune cells, such as macrophages or dendritic cells, that capture material from dead tumour cells and present it on type II MHC molecules are probably involved. Indeed, dendritic cells are required<sup>8</sup> for the maintenance of an immunotherapy response in an IFN- $\gamma$ -dependent manner.

Future studies could investigate whether immunotherapies that target both CD4 and CD8 T cells should be developed for clinical use. An obstacle to understanding and harnessing the responses of CD4 T cells for immunotherapy has long been the difficulty in identifying neoantigens that trigger such responses, as well as the need for adequate tools to monitor these responses. Alspach and colleagues' work, along with that of others<sup>9,10</sup>, suggests that this is now changing.

**Jonathan L. Linehan** and **Lélia Delamarre** are at Genentech, South San Francisco, California 94080, USA.  
e-mail: delamarre.lesia@gene.com

1. Alspach, E. *et al.* *Nature* **574**, 696–701 (2019).
2. Borst, J., Ahrends, T., Bąbala, N., Melief, C. J. M. & Kastenmüller, W. *Nature Rev. Immunol.* **18**, 635–647 (2018).
3. Tran, E. *et al.* *Science* **344**, 641–645 (2014).
4. Gubin, M. M. *et al.* *Nature* **515**, 577–581 (2014).
5. Castro, F., Cardoso, A. P., Gonçalves, R. M., Serre, K. & Oliveira, M. J. *Front. Immunol.* **9**, 847 (2018).
6. Gubin, M. M. *et al.* *Cell* **175**, 1014–1030 (2018).
7. Corthay, A. *et al.* *Immunity* **22**, 371–383 (2005).
8. Garriss, C. S. *et al.* *Immunity* **49**, 1148–1161 (2018).
9. Ott, P. A. *et al.* *Nature* **547**, 217–221 (2017).
10. Abelin, J. G. *et al.* *Immunity* <https://doi.org/10.1016/j.immuni.2019.08.012> (2019).

The authors declare competing financial interests: see [go.nature.com/33kjqo](https://go.nature.com/33kjqo) for details

This article was published online on 23 October 2019.

# Robust evidence of insect declines

William E. Kunin

Data are mounting that document widespread insect losses. A long-term research project now provides the strongest evidence of this so far, and demonstrates the value of standardized monitoring programmes. **See p.671**

There are certain times in life – whether in our relationships, personal health or scientific research – when we think that we know something but the evidence is less than conclusive. An accumulation of clues or symptoms might suggest a particular interpretation without being strong enough to clinch the argument. In such situations, it can be a relief to finally get a definitive answer, even if the news is bad. Once we know that a problem definitely exists, we may be able to do something about it. Readers might feel the same way when they read the results reported on page 671 by Seibold *et al.*<sup>1</sup>, which provide compelling evidence of a major problem – large-scale declines in the numbers and diversity of insects and other jointed invertebrates known as arthropods.

Insects have pivotal roles in terrestrial ecosystems. These organisms dominate global animal biodiversity in terms of their biomass, species numbers and total population numbers, and they perform important ecosystem functions and services such as pollinating flowers, disposing of dead organisms and waste, and forming crucial links in food webs. Insect declines have been implicated as possible drivers of declines in insect-eating birds<sup>2</sup> and in animal-pollinated plants<sup>3</sup>. Thus, massive losses in insect diversity (Fig. 1) and abundance would be grounds for serious concern.

The rumours of such declines have been around for some time. In the 1990s, researchers warned that extinctions among insects probably outstripped those of more highly studied organisms such as vertebrates and plants<sup>4</sup>. Evidence subsequently grew of declines in particular insect groups, including butterflies<sup>5</sup> and bees<sup>3</sup>. Most such studies have focused narrowly on particular insect orders or families, although a recent global meta-analysis<sup>6</sup> provides strong indications that insect losses are geographically widespread and occur across a range of taxonomic groups.

However, much of this evidence has come from biodiversity databases – records of species sightings, mostly collected by volunteers, and usually gathered in a haphazard fashion.

Analytical techniques can use such records to assess changes in the local species richness<sup>3</sup> or species' distributions<sup>7</sup>, but analyses of this sort are frustratingly indirect. The models used to analyse data can attempt to take into account the often strong temporal and spatial biases in these data sets. However, such results could still be influenced by changes in the nature of the biodiversity recording over time, fuelled by changes in observers' goals and methods. Such analyses are also limited in scope. Although they can reveal changes in diversity, such data, recording a glimpse of a species at a given site on a particular date, do not reveal shifts in insect abundance, which is arguably the most crucial aspect to assess when monitoring ecosystem services<sup>8</sup>.

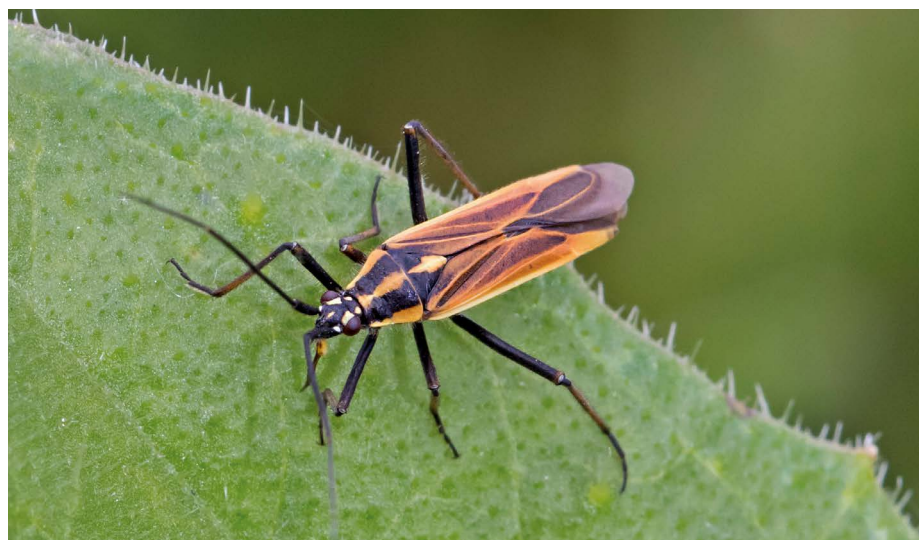
Standardized sampling can fill that gap. A previous study<sup>9</sup> reported the data collected from a network of standardized insect traps set by amateur entomologists in German nature reserves over a 27-year period. That study indicated that the biomass of insects captured declined by 75% over the period studied. This research raised serious concerns,

but it had limitations: the sampling of sites was opportunistic and not always consistent over time, and although the biomass of the specimens caught was recorded, the species were not identified or even counted, meaning that species richness and abundance couldn't be assessed.

Seibold and colleagues finally complete the circle by reporting species richness, abundance and biomass for a wide range of arthropod taxa recorded using standardized sampling<sup>1</sup>. They describe the results of monitoring over nearly ten years of intensive study in grasslands and woodlands in three regions of Germany as part of the country's interdisciplinary Biodiversity Exploratories project<sup>10</sup>.

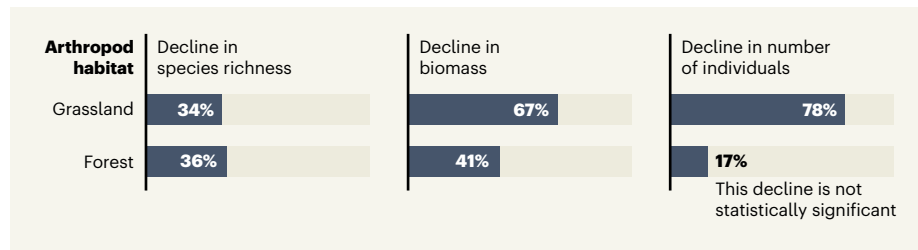
The results show clear evidence of substantial declines in arthropod abundance and biodiversity (Fig. 2). Grasslands were particularly badly affected: species richness of arthropods fell by 34% over the monitoring period, and the arthropod biomass and numbers recorded dropped by 67% and 78%, respectively. These declines were particularly strong in landscapes dominated by farmland, suggesting that agricultural management could be driving this drop. The losses among forest-dwelling arthropods were less precipitous by comparison, with a 36% drop in species richness, a 41% loss of biomass and no statistically significant population decline. The verdict is clear. In Germany at least, insect declines are real, and they're every bit as severe as had been feared.

Such long-term standardized monitoring as carried out by Seibold and colleagues is not cheap. However, the expense is dwarfed by the expenditure needed to address the problem. Agri-environmental programmes in the European Union, for example, spend tens of billions of euros to encourage farmland biodiversity, a substantial portion of which is aimed specifically or partially at insects (see go.nature.



**Figure 1 | The meadow plant bug, *Leptopterna dolabrata*.** Rumours of a decline in the numbers of insects such as the meadow plant bug are a cause for concern.





**Figure 2 | Arthropod declines recorded in Germany.** Seibold *et al.*<sup>1</sup> report nearly a decade's worth of standardized sampling of arthropods – jointed invertebrates such as insects – at grassland and forest sites. The authors provide compelling evidence of declines in arthropod populations over time. These large-scale changes will probably have a negative effect on key ecosystem services such as pollination. Seibold and colleagues found that species richness and biomass declined significantly during the course of the study in both types of habitat, as did the number of arthropod individuals recorded at grassland sites. The decline in arthropod population sizes noted at forest sites was not statistically significant.

com/35pvdv). Similarly, EU restrictions on the use of neonicotinoid insecticides were instituted specifically to protect insect pollinators, and (industry-funded) research<sup>11</sup> suggests that this restriction has cost EU oilseed-rape farmers more than €500 million (US\$549 million) annually, because of reduced production and the consequential increased costs.

Ignorance is expensive. Given that substantial investments of public and private funds for insect conservation are deemed appropriate by society, then surely it is sensible to spend a tiny proportion of such funds on monitoring, allowing us to assess the effectiveness of these

actions and to adjust them if necessary.

The Biodiversity Exploratories project would be a good model for such an effort. This is because it extends beyond population monitoring to provide a platform for cross-disciplinary science to address the large-scale and long-term issues that are crucial in driving declines of insect communities, but which aren't amenable to analysis by controlled experimentation. The project not only helps to document declines in insect populations and biodiversity, but also assists with diagnosing their potential causes. If such in-depth, landscape-scale field research and

monitoring were rolled out more widely across Europe and beyond, we could begin to build land-use and agricultural policies on the basis of compelling scientific evidence. The results reported by Seibold and colleagues might not be good news, but at least now we know where we stand and what we should start to do.

**William E. Kunin** is in the School of Biology, University of Leeds, Leeds LS2 9JT, UK.  
e-mail: w.e.kunin@leeds.ac.uk

1. Seibold, S. *et al.* *Nature* **574**, 671–674 (2019).
2. Hallmann, C. A., Foppen, R. P. B., van Turnhout, C. A. M., de Kroon, H. & Jongejans, E. *Nature* **511**, 341–343 (2014).
3. Biesmeijer, J. C. *et al.* *Science* **313**, 351–354 (2006).
4. Thomas, J. A., Morris, M. G. & Hambler, C. *Phil. Trans. R. Soc. B* **344**, 47–54 (1994).
5. Maes, D. & Van Dyck, H. *Biol. Conserv.* **99**, 263–276 (2001).
6. Sánchez-Bayo, F. & Wyckhuys, K. A. G. *Biol. Conserv.* **232**, 8–27 (2019).
7. Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P. & Roy, D. B. *Methods Ecol. Evol.* **5**, 1052–1060 (2014).
8. Winfree, R., Fox, J. W., Williams, N. M., Reilly, J. R. & Cariveau, D. P. *Ecol. Lett.* **18**, 626–635 (2015).
9. Hallmann, C. A. *et al.* *PLoS ONE* **12**, e0185809 (2017).
10. Fischer, M. *et al.* *Basic Appl. Ecol.* **11**, 473–485 (2010).
11. Noleppa, S. *et al.* *Banning neonicotinoids in the European Union*. Research paper 01/2017 (HFFA Research GmbH, 2017).



**The week's best science,  
from the world's leading  
science journal.**

**NATURE.COM/NATURE/PODCAST**

**nature**

# A 100-kiloparsec wind feeding the circumgalactic medium of a massive compact galaxy

<https://doi.org/10.1038/s41586-019-1686-1>

Received: 20 May 2019

Accepted: 16 August 2019

Published online: 30 October 2019

David S. N. Rupke<sup>1\*</sup>, Alison Coil<sup>2</sup>, James E. Geach<sup>3</sup>, Christy Tremonti<sup>4</sup>, Aleksandar M. Diamond-Stanic<sup>5</sup>, Erin R. George<sup>2</sup>, Ryan C. Hickox<sup>6</sup>, Amanda A. Kepley<sup>7</sup>, Gene Leung<sup>2</sup>, John Moustakas<sup>8</sup>, Gregory Rudnick<sup>9</sup> & Paul H. Sell<sup>4,10,11</sup>

Ninety per cent of baryons are located outside galaxies, either in the circumgalactic or intergalactic medium<sup>1,2</sup>. Theory points to galactic winds as the primary source of the enriched and massive circumgalactic medium<sup>3–6</sup>. Winds from compact starbursts have been observed to flow to distances somewhat greater than ten kiloparsecs<sup>7–10</sup>, but the circumgalactic medium typically extends beyond a hundred kiloparsecs<sup>3,4</sup>. Here we report optical integral field observations of the massive but compact galaxy SDSS J211824.06+001729.4. The oxygen [O II] lines at wavelengths of 3726 and 3729 angstroms reveal an ionized outflow spanning 80 by 100 square kiloparsecs, depositing metal-enriched gas at 10,000 kelvin through an hourglass-shaped nebula that resembles an evacuated and limb-brightened bipolar bubble. We also observe neutral gas phases at temperatures of less than 10,000 kelvin reaching distances of 20 kiloparsecs and velocities of around 1,500 kilometres per second. This multi-phase outflow is probably driven by bursts of star formation, consistent with theory<sup>11,12</sup>.

The galaxy SDSS J211824.06+001729.4 that we study here, which we call Makani (Hawai'ian for 'wind'), is an example of a merger of two galaxies hosting a galactic wind thought to be powered by extreme star-formation surface density<sup>13</sup>. At redshift  $z = 0.459$ , Makani is a compact but massive galaxy, with  $\log(M_*/M_\odot) = 11.1(\pm 0.2)$ , where  $M_*$  and  $M_\odot$  are the stellar and solar masses, respectively (Extended Data Fig. 4). Our Hubble Space Telescope imaging analysis reveals a highly peaked stellar core (radius 400 pc) framed by two tidal tails of 10–15 kpc so that half of the galaxy's light extends to about 2.5 kpc (ref. <sup>14</sup>; Fig. 1). Its stellar populations include old (more than a billion years, Gyr), medium-aged (0.4 Gyr), and young (less than 7 million years, Myr) components (Extended Data Fig. 5), with a current star-formation rate of  $100\text{--}200 M_\odot \text{ yr}^{-1}$ . It may contain a dust-obscured accreting supermassive black hole, or active galactic nucleus (AGN), on the basis of its X-ray luminosity of  $\log L(2\text{--}10 \text{ keV}) = 42.5^{+0.4}_{-0.6} \text{ erg s}^{-1}$  (ref. <sup>14</sup>), its mid-infrared slope, and the presence of highly ionized gas such as [Ne V] at wavelength  $\lambda = 3,426 \text{ \AA}$  ( $\log L = 40.6^{+0.1}_{-0.2} \text{ erg s}^{-1}$ ). However, any AGN is not currently energetically dominant<sup>13,14</sup> or radio-loud (G. C. Petter et al., manuscript in preparation) and the data could be explained by star formation and shocks. Extremely high-density star formation, like that found in Makani, is capable of powering fast winds, independent of an AGN<sup>13</sup>.

To study the spatial extent of its outflow, we observed Makani with the Keck Cosmic Web Imager (KCWI)<sup>15</sup>. The emission from the [O II] lines at  $\lambda = 3,726 \text{ \AA}$  and  $3,729 \text{ \AA}$  in these data reveals a nebula with approximate mirror symmetry around the north–south and east–west axes, extending

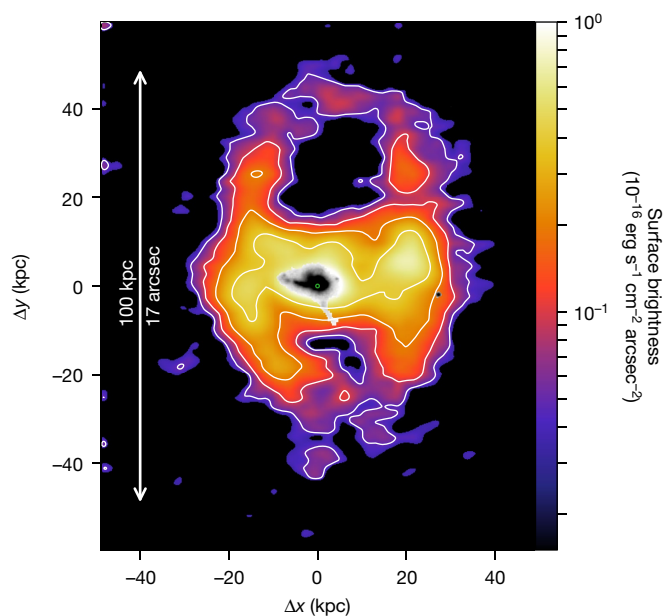
to radii of 50 kpc north and south of the galaxy nucleus, and 40 kpc east and west of it (Fig. 1). Its morphology resembles that of a limb-brightened, bipolar bubble, similar to those seen in other galactic winds<sup>16,17</sup> but on a much larger scale. This nebula is remarkable in the context of other [O II] emitters. Its area of  $4,900 \text{ kpc}^2$  ( $136 \text{ arcsec}^2$  above a  $5\sigma$  surface brightness limit per spaxel (spectral pixel) of  $5 \times 10^{-18} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ arcsec}^{-2}$ ) makes it the largest [O II] nebula detected around a single galaxy in the field<sup>18,19</sup> or in galaxy groups<sup>20,21</sup>. Its [O II] luminosity of  $3.3 \times 10^{42} \text{ erg s}^{-1}$  is several times the break in the galaxy luminosity function,  $L_*$ , at  $z \approx 0.45$  (ref. <sup>22</sup>). Its rest-frame equivalent width ( $40 \text{ \AA}$ ), half-light radius (17 kpc), and maximum radial extent (50 kpc) put it at the top end of [O II] emitters at  $z \leq 0.6$  and radio-galaxy nebulae<sup>18,23</sup>, perhaps indicative of the unusual nature of this nebula as a giant galactic wind.

On the basis of its light distribution alone, the hourglass shape of the nebula strongly suggests a bipolar galactic wind emerging from its host galaxy. The spatially resolved gas kinematics confirm this impression and separate the wind cleanly into an outer region with low-velocity gas only and an inner region containing both low- and high-velocity gas (Fig. 2). The outer region with lower velocities spans radii of 20–50 kpc, while the high-velocity gas is concentrated within a radius of about 10 kpc. We call these two wind components, and the associated starbursts that are thought to have produced them, episodes I (0.4 Gyr ago) and II (7 Myr ago). We sort spaxels by the average maximum blueshifted velocity of  $\langle v_{98\%} \rangle = -700 \text{ km s}^{-1}$ , although sorting by velocity dispersion  $\sigma$  produces similar results. Episode I then has gas with maximum

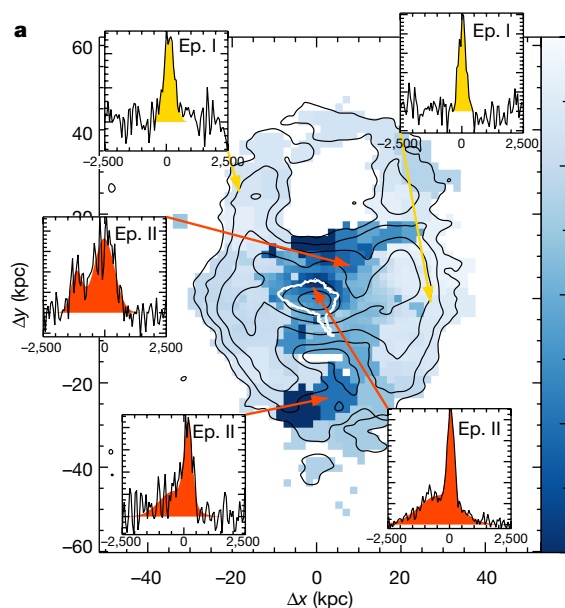
<sup>1</sup>Department of Physics, Rhodes College, Memphis, TN, USA. <sup>2</sup>Center for Astrophysics and Space Sciences, University of California, San Diego, La Jolla, CA, USA. <sup>3</sup>Centre for Astrophysics Research, School of Physics, Astronomy and Mathematics, University of Hertfordshire, Hatfield, UK. <sup>4</sup>Department of Astronomy, University of Wisconsin-Madison, Madison, WI, USA.

<sup>5</sup>Department of Physics and Astronomy, Bates College, Lewiston, ME, USA. <sup>6</sup>Department of Physics and Astronomy, Dartmouth College, Hanover, NH, USA. <sup>7</sup>National Radio Astronomy Observatory, Charlottesville, VA, USA. <sup>8</sup>Department of Physics and Astronomy, Siena College, Loudonville, NY, USA. <sup>9</sup>Department of Physics and Astronomy, University of Kansas, Lawrence, KS, USA. <sup>10</sup>Department of Astronomy, University of Florida, Gainesville, FL, USA. <sup>11</sup>University of Crete, Iraklion, Crete, Greece. \*e-mail: drupke@gmail.com

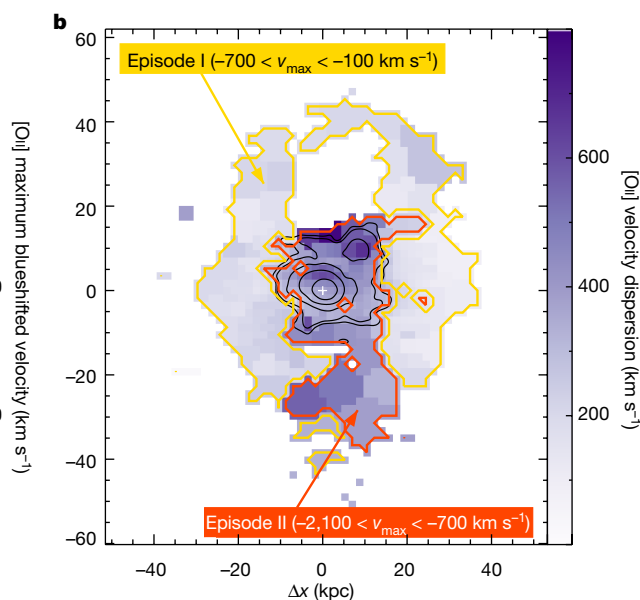




**Fig. 1 | The giant galactic wind surrounding the massive, compact galaxy Makani, observed by emission from the [O II] line at  $\lambda = 3,726 \text{ \AA}$  and  $3,729 \text{ \AA}$ .** The colour scale and white contours show observed-frame surface brightness, and the axes are labelled in kiloparsecs from the galaxy nucleus. Contours are 2–16% of peak flux, spaced by factors of 2. A rest-frame V-band image of the galaxy (Hubble Space Telescope/WFC3 F814W filter) is superimposed on the centre of the [O II] image taken with KCWI at the Keck II telescope. The small circle at the centre illustrates the radius of the compact core (400 pc). North is up and east is to the left.



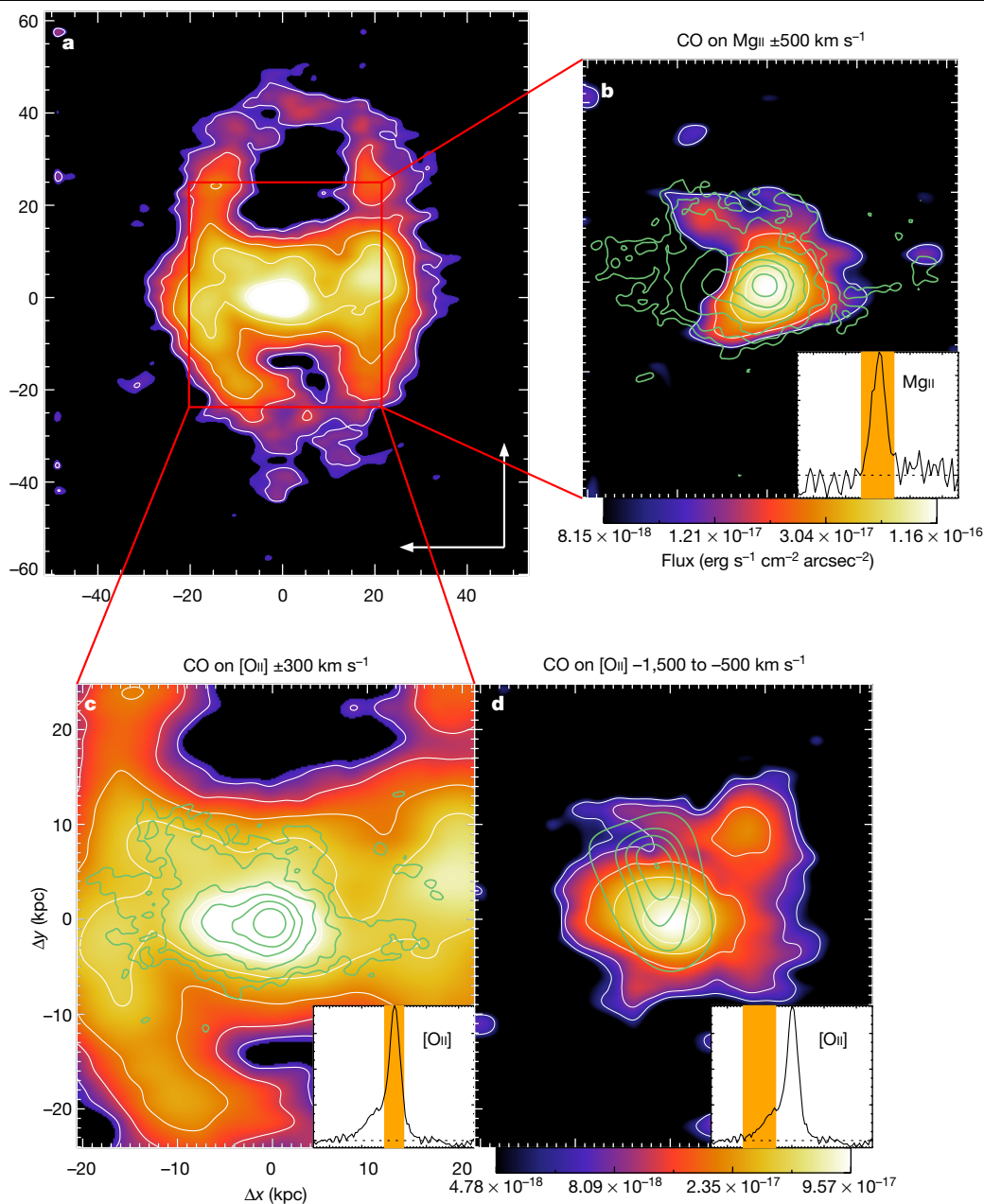
**Fig. 2 | Velocity maps of the galactic wind.** The velocities are from fits to Voronoi-binned [O II] data and are calculated from the red side of the cumulative velocity distribution function to specified percentiles as maximum blueshifted velocity  $v_{\max} \equiv v_{98\%} = v_{50\%} - 2\sigma$  (a) and velocity dispersion  $\sigma \equiv (v_{34\%} - v_{68\%})/2$  (b). In a, the blue colour denotes velocity. The insets show the deblended and stacked [O II] doublet versus velocity; these highlight representative bins for each star-formation or outflow episode. The black line is the continuum-subtracted spectrum, and the yellow or red filled profile is the emission line model. The colour denotes whether the spaxel is part of episode I (yellow) or episode II (red).



The black contours are as in Fig. 1, and the white Hubble Space Telescope contour represents 1% of the peak stellar continuum flux, accentuating the extended diffuse stellar emission from tidal forces in the merger. In b, the purple colour denotes velocity dispersion and the yellow and red contours outline regions of specified  $v_{\max} = v_{98\%}$  to delineate episodes I and II. The spatial separation of the episodes is also reflected in the  $\sigma$  map. The black contours in b outline the inner concentration of ionized gas at velocities  $-1,500$  to  $-500 \text{ km s}^{-1}$  (Fig. 3d). We securely detect high-velocity [O II] emission beyond this inner region through Voronoi binning.

blueshifted velocities between  $-100$  and  $-700 \text{ km s}^{-1}$  while episode II has a high-velocity tail of gas to  $v_{98\%} = -2,100 \text{ km s}^{-1}$ . Timescale arguments further support the existence of two starburst-driven wind episodes. In the 0.4 Gyr since the episode I starburst, a constant-velocity wind must have travelled at  $120 \text{ km s}^{-1}$  to reach the edge of the nebula (50 kpc); representative velocities at the nebula outskirts are in fact  $100$ – $200 \text{ km s}^{-1}$ . In the 7 Myr since the most recent starburst episode began, the speed required to reach the edge of the bulk of the episode II wind (10 kpc) is  $1,400 \text{ km s}^{-1}$ , which is also a typical maximum velocity in the inner nebula. The lack of high-velocity redshifted gas in episode II may be due to dust in the outflow blocking the far side of the wind, while the lack of high-velocity gas in the episode I wind is probably due to the dispersal of high-velocity gas to radii exceeding 50 kpc over 0.4 Gyr or the deceleration of the outflow as predicted by models<sup>11</sup>. Episode I gas also forms the telltale hourglass shape and has higher typical velocity dispersions ( $200 \text{ km s}^{-1}$ ) than expected for tidal features or gravitational motions at large radii<sup>24</sup>.

We find two other gas phases that we associate with the episode II (recent, inner) outflow. Using the Atacama Large Millimeter Array (ALMA), we detect molecular gas traced by CO(2–1) emission that is outflowing in a compact form, both blueshifted at  $-500$  to  $-1,500 \text{ km s}^{-1}$  and redshifted at  $500$ – $1,500 \text{ km s}^{-1}$ , from the nucleus 10 kpc northward (Fig. 3d). This gas is clearly part of episode II, given its high velocity and compact scale. Lower-velocity molecular gas, at  $|v| < 500 \text{ km s}^{-1}$  and radius  $r < 20 \text{ kpc}$  (Fig. 3c), is also likely to be part of the outflow because it is much more extended than the stellar disk and correlates spatially with extended, outflowing ionized gas. It may be gas from episode II that has decelerated after reaching scales of the order of 10 kpc. Using resonant line emission from Mg II at  $\lambda = 2,796 \text{ \AA}$  and  $2,803 \text{ \AA}$ , we also detect neutral gas of temperature  $T \approx 10^4 \text{ K}$  in the velocity range  $\pm 500 \text{ km s}^{-1}$  (Fig. 3b). This emission correlates with some regions of faint, extended CO and [O II] emission. Although these velocities are modest, resonant emission on 10-kpc scales has so far been detected



**Fig. 3 | Comparison of the ionized, neutral atomic and molecular phases of the galactic wind.** The [O II] map from Fig. 1 is replicated in **a**, with the north and east directions indicated in the lower right. In **b**, a zoomed-in view of the inner 40 kpc, molecular gas in restored CO(2–1) (green contours) is plotted on top of Mg II emission (colour, with white contours) in the same velocity range (–500 to +500 km s<sup>–1</sup>). Panel **c** compares the low-velocity molecular gas (restored CO, green contours) and ionized gas ([O II], colour and white contours) over the

same velocities (–300 to +300 km s<sup>–1</sup>). The CO contours are 0.09, 0.14, 0.28, 0.5, 0.8 and 1.2 mJy per beam. Panel **d** is the high-velocity molecular gas (tapered, bottom right, contours of 0.08, 0.11, 0.122, 0.132 and 0.14 mJy per beam) on an [O II] channel map showing the brightest regions of high-velocity ionized gas, both over –1,500 to –500 km s<sup>–1</sup>. In each zoom box, the inset spectrum is a plot of spatially integrated optical line flux versus velocity (–2,500 to +2,500 km s<sup>–1</sup>), with the linemap velocity range highlighted in orange.

only in galactic winds, and because of strong radiation transfer effects such emission is not highly shifted from the redshift of the Makani galaxy’s centre of mass<sup>25,26</sup>. Blueshifted Fe II absorption is detected in the nuclear spectrum (Extended Data Fig. 2), but tracing its physical extent requires deeper observations.

Estimates of the mass contained within the wind would complete its portrait. However, the mass of the ionized gas is uncertain without spatially resolved recombination line measurements. Bootstrapping from single-aperture H $\alpha$  and H $\beta$  measurements, we estimate  $6_{-3}^{+6} \times 10^8 (200 \text{ cm}^{-3}/n_e) M_\odot$  of ionized gas in the nebula, with unquantified systematic errors (electron density  $n_e$  and ionization state) probably exceeding the measurement error. Although it is confined to the inner

10 kpc, the mass of the molecular gas in the  $|v| = 500\text{--}1,500 \text{ km s}^{-1}$  flow (that is, episode II) is substantial ( $2.4_{-0.6}^{+0.9} \times 10^9 M_\odot$ ), with four times as much in the more extended  $\pm 500 \text{ km s}^{-1}$  component. The ionized gas plus molecular wind thus contains 1–10% of the galaxy’s baryonic mass, a fraction that will be even larger when all phases are accounted for. The resulting mass flow rate for the molecular episode II component is  $dM/dt \approx Mv/r = 245 M_\odot \text{ yr}^{-1}$  for  $v = 1,000 \text{ km s}^{-1}$  and  $r = 10 \text{ kpc}$ , which is roughly one to two times the star-formation rate. This is consistent with molecular outflow rates<sup>710</sup> from other compact starburst mergers at  $z > 0.5$ .

The huge, metal-enriched outflows in Makani—a key component of the host galaxy’s dynamically and chemically evolving circumgalactic medium (CGM)—are consistent with the types of star-formation



or AGN-driven winds that populate and enrich the CGM in theoretical models<sup>11,12</sup>. A model galaxy forming stars at  $100M_{\odot} \text{ yr}^{-1}$  in a baryonic halo of  $2 \times 10^{11}M_{\odot}$  and supernova-driven winds propelling gas with initial velocity  $1,000 \text{ km s}^{-1}$  produce a  $10^9M_{\odot}$  (or  $10^{10}M_{\odot}$ ) shell at  $r \approx 10 \text{ kpc}$  (or  $100 \text{ kpc}$ ) in  $t \approx 10 \text{ Myr}$  (or  $400 \text{ Myr}$ ), with velocities at this time of the order of  $1,000 \text{ km s}^{-1}$  ( $100 \text{ km s}^{-1}$ )<sup>11</sup>. These numbers bear a striking resemblance to the observations in the context of the two-episode outflow we propose. The wind will continue to expand, diffuse, and virialize over longer timescales, as this nebula is denser and more structured than virialized CGM gas<sup>4,12</sup> and has reached perhaps only about 10% of the virial radius<sup>27</sup> of a galaxy with  $\log(M_{*}/M_{\odot}) = 11.1$  at  $z = 0.46$ .

The size of this wind makes it the one of the largest wide-angle, galaxy-scale outflows yet observed, with scales much larger than in other compact or high- $z$  starbursts<sup>7,9,10</sup>. The morphology and velocity of the wind in the Teacup AGN make it a cousin of Makani, but on scales five to ten times smaller<sup>17</sup>. Although the Teacup also hosts diffuse gas over 100-kpc scales, this gas has a different physical origin than does the gas in Makani<sup>28</sup>. The most comparable system may be a merger with compact star formation at a cosmic distance ten times closer than Makani. NGC 6240 ( $z = 0.04$ ) has an ionized outflow that reaches a 40-kpc radius<sup>8</sup>. However, the outflow size relative to the stellar half-light radius (inside which half of the galaxy's starlight resides) is only  $r/r_{*,1/2} \approx 4$  in NGC 6240 (ref. <sup>29</sup>), versus  $r/r_{*,1/2} > 20$  in Makani, and much of the NGC 6240 nebula coincides with stellar tidal features, unlike Makani. Furthermore, the H $\alpha$  luminosity of the NGC 6240 nebula is four times smaller than in Makani if the core H $\alpha$  emission follows the oxygen emission at a constant [O II]/H $\alpha$  line ratio.

With a maximum extent of more than twenty times the stellar half-light radius, the oxygen nebula observed here has propagated well into the galaxy halo, placing it solidly in the CGM. The cool gas and metals in the flow are thus contributing to the buildup and enrichment of the CGM. This cool gas can be propelled by hot gas, radiation pressure or cosmic rays. The classic model of hot gas acceleration faces the problem that cold clouds may be destroyed during acceleration. These clouds may simply reform after being shredded and mixed in the hot wind<sup>30</sup> or the clouds may survive the acceleration through fast radiative cooling<sup>31</sup>. The mixing layers in shredded clouds can also cool hot gas from the halo or CGM, enhancing the amount of cool gas injected into the CGM<sup>31</sup>. If the cool outflow is accelerated by a hot wind, the existence of [O II]-emitting gas at all radii argues for either cloud reformation on very short timescales or for cloud survival (coupled with enhancement from hot gas). The outflow we observe is thus feeding the CGM by directly depositing gas from the galaxy or by entraining and cooling hot halo and circumgalactic gas.

Connecting the CGM with ongoing galactic winds has been challenging because of the lack of clear evidence for such winds on large enough scales. Previous evidence came from theory<sup>5,6</sup> and the statistical characteristics of the CGM as measured from single quasar absorption lines over large galaxy populations<sup>3,4</sup>. We have now observed a single galaxy, with all lines of sight accounted for, whose wind has entered the CGM. Our measurement provides one of the first direct windows into the dynamically and chemically evolving, multiphase CGM being created around a massive galaxy.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1686-1>.

- Shull, J. M., Smith, B. D. & Danforth, C. W. The baryon census in a multiphase intergalactic medium: 30% of the baryons may still be missing. *Astrophys. J.* **759**, 23 (2012).
- Tumlinson, J., Peebles, M. S. & Werk, J. K. The circumgalactic medium. *Annu. Rev. Astron. Astrophys.* **55**, 389–432 (2017).
- Tumlinson, J. et al. The large, oxygen-rich halos of star-forming galaxies are a major reservoir of galactic metals. *Science* **334**, 948–952 (2011).
- Werk, J. K. et al. The COS-Halos survey: physical conditions and baryonic mass in the low-redshift circumgalactic medium. *Astrophys. J.* **792**, 8 (2014).
- Hummels, C. B., Bryan, G. L., Smith, B. D. & Turk, M. J. Constraints on hydrodynamical subgrid models from quasar absorption line studies of the simulated circumgalactic medium. *Mon. Not. R. Astron. Soc.* **430**, 1548–1565 (2013).
- Ford, A. B. et al. Hydrogen and metal line absorption around low-redshift galaxies in cosmological hydrodynamic simulations. *Mon. Not. R. Astron. Soc.* **432**, 89–112 (2013).
- Geach, J. E. et al. Stellar feedback as the origin of an extended molecular outflow in a starburst galaxy. *Nature* **516**, 68–70 (2014).
- Yoshida, M. et al. Giant H $\alpha$  nebula surrounding the starburst merger NGC 6240. *Astrophys. J.* **820**, 48 (2016).
- Falgarone, E. et al. Large turbulent reservoirs of cold molecular gas around high-redshift starburst galaxies. *Nature* **548**, 430–433 (2017).
- Geach, J. E. et al. Violent quenching: molecular gas blown to  $1000 \text{ km s}^{-1}$  during a major merger. *Astrophys. J.* **864**, L1 (2018).
- Lochhaas, C., Thompson, T. A., Quataert, E. & Weinberg, D. H. Fast winds drive slow shells: a model for the circumgalactic medium as galactic wind-driven bubbles. *Mon. Not. R. Astron. Soc.* **481**, 1873–1896 (2018).
- Hani, M. H., Sparre, M., Ellison, S. L., Torrey, P. & Vogelsberger, M. Galaxy mergers moulding the circumgalactic medium I. The impact of a major merger. *Mon. Not. R. Astron. Soc.* **475**, 1160–1176 (2018).
- Diamond-Stanic, A. M. et al. High-velocity outflows without AGN feedback: Eddington-limited star formation in compact massive galaxies. *Astrophys. J.* **755**, L26 (2012).
- Sell, P. H. et al. Massive compact galaxies with high-velocity outflows: morphological analysis and constraints on AGN activity. *Mon. Not. R. Astron. Soc.* **441**, 3417–3443 (2014).
- Morrissey, P. et al. The Keck Cosmic Web Imager Integral Field Spectrograph. *Astrophys. J.* **864**, 93 (2018).
- Shopbell, P. L. & Bland-Hawthorn, J. The asymmetric wind in M82. *Astrophys. J.* **493**, 129–153 (1998).
- Harrison, C. M. et al. Storm in a “Teacup”: a radio-quiet quasar with  $\approx 10 \text{ kpc}$  radio-emitting bubbles and extreme gas kinematics. *Astrophys. J.* **800**, 45 (2015).
- Bridge, J. S. et al. Physical and morphological properties of [O II] emitting galaxies in the HETDEX pilot Survey. *Astrophys. J.* **799**, 205 (2015).
- Yuma, S. et al. Systematic Survey for [O II], [O III], and H $\alpha$  blobs at  $z = 0.1$ – $1.5$ : the implication for evolution of galactic-scale outflow. *Astrophys. J.* **841**, 93 (2017).
- Epinat, B. et al. Ionised gas structure of 100 kpc in an over-dense region of the galaxy group COSMOS-Gr30 at  $z \sim 0.7$ . *Astron. Astrophys.* **609**, A40 (2018).
- Johnson, S. D. et al. Galaxy and quasar fueling caught in the act from the intragroup to the interstellar medium. *Astrophys. J.* **869**, L1 (2018).
- Zhu, G., Moustakas, J. & Blanton, M. R. The [O II]  $\lambda 3727$  luminosity function at  $z \sim 1$ . *Astrophys. J.* **701**, 86–93 (2009).
- Kalfountzou, E., Jarvis, M. J., Bonfield, D. G. & Hardcastle, M. J. Star formation in high-redshift quasars: excess [O II] emission in the radio-loud population. *Mon. Not. R. Astron. Soc.* **427**, 2401–2410 (2012).
- Pulsoni, C. et al. The extended Planetary Nebula Spectrograph (eP.N.S.) early-type galaxy survey: the kinematic diversity of stellar halos and the relation between halo transition scale and stellar mass. *Astron. Astrophys.* **618**, A94 (2018).
- Rubin, K. H. R. et al. Low-ionization line emission from a starburst galaxy: a new probe of a galactic-scale outflow. *Astrophys. J.* **728**, 55 (2011).
- Prochaska, J. X., Kasen, D. & Rubin, K. Simple models of metal-line absorption and emission from cool gas outflows. *Astrophys. J.* **734**, 24 (2011).
- Behroozi, P. S., Conroy, C. & Wechsler, R. H. A comprehensive analysis of uncertainties affecting the stellar mass-halo mass relation for  $0 < z < 4$ . *Astrophys. J.* **717**, 379–403 (2010).
- Villar-Martín, M. et al. A 100 kpc nebula associated with the ‘Teacup’ fading quasar. *Mon. Not. R. Astron. Soc.* **474**, 2302–2312 (2018).
- Kim, D. C. et al. Hubble Space Telescope ACS imaging of the GOALS sample: quantitative structural properties of nearby luminous infrared galaxies with  $L_{\text{IR}} > 10^{11.4} L_{\odot}$ . *Astrophys. J.* **768**, 102 (2013).
- Thompson, T. A., Quataert, E., Zhang, D. & Weinberg, D. H. An origin for multiphase gas in galactic winds and haloes. *Mon. Not. R. Astron. Soc.* **455**, 1830–1844 (2016).
- Gronke, M. & Oh, S. P. The growth and entrainment of cold gas in a hot wind. *Mon. Not. R. Astron. Soc.* **480**, L111–L115 (2018).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### KCWI observations and data analysis

SDSS J211824.06+001729.4 was originally selected as an intermediate-redshift starburst galaxy with broad but spatially unresolved line emission<sup>13,14</sup>, part of a population known to host strong outflows<sup>32</sup>. We observed it with KCWI on the Keck II telescope on 6 November 2018 UT (Universal Time) for 40 min. We used the blue low-dispersion (BL) grating and medium slicer with KCWI, yielding a resolution of 2.5 Å and wavelength coverage of 3,435–5,525 Å. We chose a central wavelength of 4,500 Å and detector binning of  $2 \times 2$ . Conditions were photometric, with 0.6'' seeing. Two exposures were dithered 0.35'' along slices to subsample the long spatial dimension of the output spaxels ( $0.69'' \times 0.29''$ ). The field of view of the reduced data cube is  $15.4'' \times 19.4''$ .

We reduced the data using the KCWI data reduction pipeline and the IFSRED library<sup>33</sup>. As the default scattered light subtraction in the pipeline leaves visible residuals, we use a routine (IFSR\_KCWISCAT-SUB) to subtract scattered light by summing the data in 100-pixel increments along columns (parallel to the dispersion direction) and fitting the least-contaminated inter-slice regions along rows (parallel to the spatial direction) with low-order polynomials. The default wavelength calibration also produces large root-mean-square (r.m.s.) residuals because of a mismatch with the pipeline thorium-argon (ThAr) atlas, so we extract a representative spectrum from our data and find its wavelength solution using IDENTIFY in PyRAF ([www.stsci.edu/institute/software\\_hardware/pyraf](http://www.stsci.edu/institute/software_hardware/pyraf)). 16/20 lines in the 3,500–4,000 Å and 5,000–5,600 Å ranges are from Th; the other 38 lines (including most of the brightest lines) are from Ar. The resulting r.m.s. residual is 0.18 Å. We then input this calibrated spectrum as the atlas into the pipeline, yielding a 0.07 Å r.m.s. Following the pipeline stages, we resample the data (IFSR\_KCWIRESAMPLE) onto a  $0.29'' \times 0.29''$  spaxel grid; align the two exposures by fitting the galaxy centroid (IFSR\_PEAK); and mosaic the data (IFSR\_MOSAIC). The resulting stacked and resampled field of view at 5,000 Å is  $53 \times 67$  spaxels. The reconstructed KCWI continuum image (rest-frame near-ultraviolet) is consistent with the Hubble Space Telescope WFC3/F814W image (rest-frame *I*) when convolved with a 15-pixel Gaussian kernel to match the measured seeing. Finally, we sum the nebula's core emission in a  $3.0''$  circular aperture to match the Sloan Digital Sky Survey (SDSS)<sup>34</sup> spectrum.

We created initial [O II] linemaps by integrating over [O II] and subtracting nearby continuum windows on either side. The wavelength interval of each map is calculated from the doublet average wavelength at a given velocity. We then used  $\pm 300 \text{ km s}^{-1}$  flux and error maps to create Voronoi bins (where the velocity applies to the centroid of the [O II] doublet; the doublet lines are 2.6 Å apart, which corresponds to  $200 \text{ km s}^{-1}$ ). The IDL ([www.harrisgeospatial.com/Software-Technology/IDL](http://www.harrisgeospatial.com/Software-Technology/IDL)) routine VORONOI\_2D\_BINNING<sup>35</sup> is used to construct the bins, with a target signal-to-noise ratio of 10 and a threshold signal-to-noise ratio of 1. We fit the core spectrum, the full data cube, and the Voronoi binned data cube with IFSFIT<sup>36</sup>. Because very few strong stellar lines arise in our spectra (rest-frame 2,350–3,790 Å), we use a scaled continuum derived from the fit to the rest-frame 2,550–5,600 Å spectrum<sup>14</sup> (see below). We fit two velocity components to the [O II] and [Ne V] lines. If any component falls below  $2\sigma$  in a spaxel, the spectrum is re-fitted with fewer components. Allowing the [O II] line ratio to float freely in the narrow component of the core spectrum results in an [O II] 3,729 Å / [O II] 3,726 Å ratio of 1.2 (corresponding to  $n_e \approx 200 \text{ cm}^{-3}$ ), while the broad component ratio is unconstrained. We thus fix the [O II] ratio to 1.2 in all fits. In the core spectrum, Mg I 2,852 Å, Mg II 2,796 Å, 2,803 Å, and Fe II\* 2,612 Å, 2,626 Å are tied to the same velocity and width and fitted with a single component. The continuum fits to each spaxel are used to subtract the stellar continuum around [O II] or Mg II to produce the linemaps shown in Figs. 1–3. The [O II] linemaps have a limiting  $1\sigma$  surface brightness per pixel of  $1.0 \times 10^{-18} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ arcsec}^{-2}$ . For display

purposes only, these maps are interpolated to a grid ten times finer and the  $\pm 300 \text{ km s}^{-1}$  maps are clipped at 1.5% of peak flux (or  $4\sigma$ ).

The core spectrum yields detections of [O II], [Ne V] 3,426 Å, Mg I 2,852 Å, Mg II 2,796 Å, 2,803 Å, and Fe II\* 2,612 Å, 2,626 Å in emission, and Fe II 2,586 Å in absorption. The emission lines break into two distinct components: a narrow feature at the systemic velocity of the host galaxy ( $z = 0.45916$ ;  $\sigma = 143 \text{ km s}^{-1}$  in [O II],  $197 \text{ km s}^{-1}$  in Mg and Fe emission) and a broad, blueshifted feature that is outflowing ( $z = 0.45736$ ,  $v = -540 \text{ km s}^{-1}$ ,  $\sigma = 500 \text{ km s}^{-1}$  in [O II];  $z = 0.45666$ ,  $v = -750 \text{ km s}^{-1}$ ,  $\sigma = 392 \text{ km s}^{-1}$  in [Ne V]). We correct upward spatially integrated line fluxes and luminosities for a Galactic extinction of  $A_V = 0.2075$  (ref. <sup>37</sup>), which corresponds to a 22% correction at [O II]. We use the 2018 Planck cosmology<sup>38</sup> to calculate luminosity and angular size distances. The spatially integrated [O II] flux is  $4.0 (\pm 0.2) \times 10^{-15} \text{ erg s}^{-1} \text{ cm}^{-2}$ . From the core spectrum we measure a Mg II flux of  $2.3 (\pm 0.1) \times 10^{-16} \text{ erg s}^{-1} \text{ cm}^{-2}$ , which corresponds to a rest-frame equivalent width of 2.3 Å and a luminosity of  $1.5 \times 10^{41} \text{ erg s}^{-1}$ . The [Ne V] line flux is  $4.3 (\pm 1.2) \times 10^{-17} \text{ erg s}^{-1} \text{ cm}^{-2}$ .

### Ionized and neutral gas properties

We parameterize the velocity distribution of the [O II]-emitting gas in each spaxel using the cumulative velocity distribution function (CVDF). In spaxels where only one component is fitted, this is a Gaussian; for two components, the CVDF is the sum of two independent Gaussians. We use the 50th and 98th percentiles of the CVDF ( $v_{50\%}$  and  $v_{98\%}$ , as measured from the red side of the line) to represent the mean and most blueshifted ('maximum') velocities. We define the width of the CVDF as  $\sigma = (v_{34\%} - v_{68\%})/2$ , which for a single component is the usual Gaussian  $\sigma$ .

We combine our KCWI data with two other spectra to constrain the integrated gas excitation, reddening and gas mass. The first is the SDSS spectrum that, along with [O II], covers the [Ne III] 3,869 Å, H $\beta$ , and [O III] 4,959 Å, 5,007 Å emission lines. The second is a spectrum acquired with Keck/NIRSPEC, which covers the H $\alpha$ , [N II] 6,548 Å, 6,583 Å, and [S II] 6,717 Å, 6,731 Å lines. The latter was observed with a  $0.76''$ -wide slit at a position angle of  $83^\circ$  east of north. We scale the NIRSPEC data to match the SDSS spectrum where they overlap and correct for Galactic extinction as above. Because of the lower signal-to-noise ratio in these spectra compared to the KCWI spectrum, we fix the velocities and linewidths of each emission line using the fit to the KCWI core spectrum. We measure an H $\alpha$  flux of  $1.9 (\pm 0.1) \times 10^{-15} \text{ erg s}^{-1} \text{ cm}^{-2}$  and an extinction of  $E(B-V) = 0.4 (\pm 0.2)$  from the Balmer decrement. We then scale this flux upward to account for the entire nebula, since the [O II] flux within the SDSS aperture is 24% of the total. Using the estimated gas density of  $200 \text{ cm}^{-3}$  from the [O II] flux ratio and correcting upward for extinction by a factor of four yields an ionized gas mass of  $6^{+6}_{-3} \times 10^8 (200 \text{ cm}^{-3}/n_e) M_\odot$ . The extinction uncertainty drives the 50% error, but unquantified systematic uncertainties in the electron density and [O II]/H $\alpha$  line ratio (because we do not spatially resolve these quantities) are likely to be larger than this.

We show rest-frame optical line flux ratios from the unresolved NIRSPEC and SDSS spectra in Extended Data Fig. 1. The narrow component is consistent with photo-ionization by young stars and a near-solar metallicity, while the broad component is consistent with photo-ionization by an AGN<sup>39</sup> with ionization parameter  $U \approx -2$  or shock photo-ionization<sup>40</sup> with a velocity of at least  $300\text{--}400 \text{ km s}^{-1}$ . The higher excitation of the broad, outflowing component is also illustrated in the increasing [O III]/[O II] ratio with increasing blueshift (Extended Data Fig. 2). Besides the ionized gas lines, Extended Data Fig. 2 shows the absorption-line outflow in Fe II 2,586 Å; the corresponding Mg II emission that is systemic with a slight red wing (typical for the emission component of a resonant-line profile in a neutral outflow but without the usual absorption)<sup>25,26</sup>; and the broad, high-velocity wings of the CO(2-1) profile.

Other high-ionization lines arise in the observed-frame optical part of the spectrum. Notably, the [Ne V] emission is spatially unresolved and found only in the outflowing component, with  $v_{98\%} = 1,500 \text{ km s}^{-1}$ . Whereas [Ne V] is typically used as an AGN indicator<sup>41</sup>, its (extincted)



# Article

luminosity in Makani,  $3.6(\pm 1.0) \times 10^{40}$  erg s<sup>-1</sup>, is three times lower than the average for typical [Ne v] emitters detected at  $z = 0.6\text{--}1.2$  (ref. <sup>42</sup>); it may therefore be emitted in shocks<sup>40,43</sup>. The line ratios in the broad component of the core KCWI and SDSS spectra of  $\log([\text{Ne v}]/[\text{Ne III}])$   $3,869 \text{ \AA}$ ) =  $-0.77^{+0.13}_{-0.20}$ ,  $\log([\text{O II}]/[\text{O III}])$  =  $-0.11^{+0.04}_{-0.05}$ ,  $\log([\text{Ne v}]/[\text{O II}])$  =  $-1.07^{+0.11}_{-0.15}$  and  $\log(\text{He II } 4,686 \text{ \AA}/\text{H}\beta)$  <  $-0.87^{+0.39}_{-0.39}$  are also consistent with either AGN photo-ionization<sup>39</sup> or ionization in shocks with velocities<sup>40,44,45</sup> of at least 300–400 km s<sup>-1</sup>.

## ALMA observations, data reduction and analysis

Makani was observed by the ALMA 12-m array as part of projects 2016.1.01072.S and 2017.1.01318.S on 11 March 2017, 10 April 2018 and 15 December 2017 in antenna configurations C40-1 (baselines 15–287 m) and C43-3 (baselines 15–500 m) and C43-6 (baselines 15–2,517 m) respectively. We used the Band 4 receivers with a representative frequency of 158.01 GHz to detect CO(2–1) at the redshift of the target. The total integration time on source was 212 min. The average precipitable water vapour column during observations was approximately 2.5 mm and the average system temperature was approximately 75 K. The atmospheric, bandpass, pointing, phase and flux calibrators included the sources J2148+0657, J2134-0153 and Neptune.

We use the quality-checked ALMA pipeline-calibrated products, concatenating the observations into a single measurement set. We image the data using CASA (version 5.1.0-74), producing three versions of the data cube: naturally weighted, 1'' tapered and 0.6'' restored. The latter uses a circular Gaussian restoring beam with a full-width at half-maximum (FWHM) of 0.6'' to match the seeing of the KCWI data. We produce a tapered image in order to maximize sensitivity to potentially extended but weak CO emission around the target. We produce clean cubes by first generating dirty cubes and assess the r.m.s. noise per channel in each version. This value is then used in an iterative cleaning step (CASA clean) where we set a cleaning threshold of  $3\sigma$ , chosen to maintain a balance between producing a clean image while ensuring that real faint extended structure is not removed. We use multi-scale cleaning with scales of 0'', 0.4'', 0.8'' and 1.6''. The FWHM of the synthesized clean beams in the naturally weighted and tapered images are  $0.39'' \times 0.31''$  position angle PA =  $-73^\circ$  east of north) and  $1.15'' \times 1.08''$  (PA =  $-79^\circ$ ) respectively. The beam-restored image by definition has a circular beam of FWHM = 0.6''. We produce data cubes with a spectral resolution of 30 km s<sup>-1</sup> (16 MHz), and image the full spectral coverage including basebands placed to measure continuum emission at 2 mm in line-free regions. The r.m.s. (1 $\sigma$ ) noise per 30 km s<sup>-1</sup> channel in the natural, tapered and restored cubes is 0.13 mJy beam<sup>-1</sup>, 0.20 mJy beam<sup>-1</sup> and 0.15 mJy beam<sup>-1</sup> respectively.

After examining the cubes and extracting spectra, we detect a weak 2 mm continuum component to the observed emission: averaged over 143.5–146.5 GHz, with a total flux density  $S_{2\text{mm}} = 26(\pm 10)$  mJy. By collapsing the cube over this frequency range we construct a continuum image that is subtracted from the channels spanning the CO(2–1) line. CO(2–1) emission is observed out to a high velocity of  $v = \pm 1,500$  km s<sup>-1</sup> in the total spectrum, similar to the maximum velocities in the [O II] nebula. Therefore, to measure the line luminosity we first average the tapered cube over  $v = \pm 1,500$  km s<sup>-1</sup> and define a  $3\sigma$  mask for the source extent, based on the noise in the channel-averaged map. This mask is used to integrate the spectrum, measuring  $\Delta V$  across different velocity ranges. We define a second mask where regions with signal exceeding 15 $\sigma$  in the collapsed tapered image are excluded, eliminating the contribution from the bright core and eastern tidal arm. The rationale for this is to provide an estimate of the CO emission associated with extended (possibly outflowing) material.

Line luminosities are calculated in the conventional radio units of K km s<sup>-1</sup> pc<sup>2</sup> as  $L' = 3.25 \times 10^7 D_L^2 (1+z)^{-1} v_{\text{rest}}^{-2} \Delta V$ , where  $D_L$  is the luminosity distance in Mpc,  $v$  is the rest-frame frequency of the line in GHz and  $\Delta V$  is the integrated line flux in Jy km s<sup>-1</sup>. CO line luminosities are converted to estimates of the molecular gas mass using  $M_{\text{H}_2} = \alpha L'$ . We quote gas masses including the highest signal-to-noise features, but masking them lowers

these masses by factors of <2. As in previous work<sup>7</sup>, we adopt  $\alpha = 0.34 M_\odot$  (K km s<sup>-1</sup> pc<sup>2</sup>)<sup>-1</sup>, which is lower than both the standard Galactic and ULIRG conversions because high-velocity extended and/or CO emission might be optically thin if it is tracing molecular gas in a turbulent outflow<sup>46</sup>. This provides a conservative estimate of the molecular gas mass.

## Size measurements

A Sérsic fit to the Hubble Space Telescope image of Makani yields an effective radius  $R_e = 2.24$  kpc for a Sérsic index of  $n = 4^{14}$ . For a pure Sérsic profile,  $R_e$  is equivalent to the stellar half-light radius  $r_{*,1/2}$ , or the radius within which half of the stellar light arises<sup>47</sup>. The substantial extended, asymmetric tidal structure in a merger like Makani will affect the determination of any Sérsic component, although in this case it appears not to be a large effect; a direct measure of the half-light radius from integration of the stellar light yields  $r_{*,1/2} \approx 2.75$  kpc. We take the average of these estimates, 2.5 kpc, to be the half-light radius.

Makani has a peaked core that is well interior of the half-light radius. An estimate of its size is the radial width at half-maximum of the radial light profile. This measure yields a core radius of 400 pc, within which 10% of the galaxy's stellar light resides. This radius is comparable to other starbursts and post-starbursts without extended tidal structure<sup>14</sup>.

For comparison, Extended Data Fig. 3 shows the radial profile of the [O II] nebula, determined from azimuthal averages over pixels in bins of radial width 2 kpc. Integrating over the nebula from the centre outward as a fraction of the total flux within 50 kpc yields a half-light radius in [O II] of 17 kpc. The short (east-to-west) and long (north-to-south) axis profiles, averaged in the direction perpendicular to each profile over bins 2 kpc wide, decrease less steeply, with maximum nuclear distances of about 40 kpc and 50 kpc along the short and long axes, respectively. When doubled, these yield the quoted size of 100 kpc  $\times$  80 kpc. These measurements approach the size of the KCWI field of view, from which we infer that the nebula could be larger.

## Stellar mass estimation

We estimate the stellar mass of Makani using the Bayesian stellar population synthesis modelling code Prospector<sup>48</sup> and the Flexible Stellar Population Synthesis (FSPS)<sup>49,50</sup> models (Extended Data Fig. 4). We assemble the spectral energy distribution at rest-frame wavelengths between 0.1  $\mu\text{m}$  and 15  $\mu\text{m}$  from the Galaxy Evolution Explorer<sup>51</sup>, the SDSS<sup>34</sup>, the Spitzer Space Telescope<sup>52</sup> and the Wide-field Infrared Survey Explorer (WISE)<sup>53</sup>. We adopt a Salpeter initial mass function from the range (0.1–100)  $M_\odot$  and assume a 'delayed  $\tau$ ' backbone star-formation history ( $\tau$  is the e-folding star-formation timescale) with a late-time burst of star formation superimposed. We assume a power-law dust attenuation curve (proportional to  $\lambda^{-0.7}$ ) and allow differential attenuation between the light from young stars relative to the diffuse interstellar medium<sup>54</sup>. Finally, we compute the infrared spectrum using energy balance arguments and basic assumptions about the re-radiated infrared spectrum<sup>55</sup>. The median value of the marginalized posterior probability for stellar mass is  $\log(M_*/M_\odot) = 11.07$  with an interquartile range of 10.98–11.14. To account for systematic uncertainties in the star formation history and other prior parameters we adopt an average stellar mass and uncertainty of  $\log(M_*/M_\odot) = 11.1(\pm 0.2)$ .

We assume that the mid-infrared dust emission in Makani arises from star formation in order to fit the spectral energy distribution. WISE mid-infrared colours  $W1 - W2 = 0.74(\pm 0.03)$  and  $W2 - W3 = 3.64(\pm 0.08)$ , in Vega magnitudes<sup>53,56</sup>—place this galaxy in a region occupied partly by starbursts, but also characteristic of obscured AGN in merging galaxies<sup>57,58</sup>. The present data do not distinguish between these possibilities.

## Stellar continuum modelling

To obtain the best constraints on the young stellar populations in Makani we fit its rest-frame ultraviolet–optical spectrum with stellar population synthesis models. This fitting is very sensitive to both the quality of the spectrophotometry and the strong stellar absorption

lines in the 3,700–5,000 Å range. Since the KCWI spectrum does not extend redwards of rest-frame 3,800 Å, we use a spectrum obtained with the Blue Channel Spectrograph on the MMT with a 1" slit<sup>44</sup>. To further extend the wavelength coverage, we join the MMT and SDSS spectra near 4,600 Å. These spectra have similar spectral resolutions ( $R \approx 1,500$ ). The combined spectrum (Extended Data Fig. 5) matches the SDSS *ugriz* photometry well, indicating good spectrophotometric calibration.

We fit the MMT+SDSS spectrum with a combination of simple stellar population models and the Salim attenuation curve<sup>59</sup>. We use FSPS to generate simple stellar populations with Padova 2008 isochrones, a Salpeter initial mass function, and a new theoretical stellar library C3K (C. Conroy et al., manuscript in preparation) with a resolution of  $R \approx 10,000$ . We use solar-metallicity simple stellar population templates with 42 ages spanning 1 Myr to 7.9 Gyr. We perform the fit with the Penalized Pixel-Fitting (pPXF) code<sup>60,61</sup>. Because the galaxy is very compact and much of its dust is likely to be in the outflow, we require all stellar populations to share the same attenuation. The best-fitting model has  $z = 0.4590$ , a stellar velocity dispersion  $\sigma = 170 \text{ km s}^{-1}$ , and  $E(B - V) = 0.19$ . The spectrum is dominated by a mixture of young and intermediate-age stellar populations, with approximately 50% of the continuum emission at rest-frame 5,500 Å contributed by populations less than 7 Myr old. An additional 40% comes from a 0.4-Gyr-old stellar population. This implies two major starburst episodes, with the 0.4-Gyr burst perhaps corresponding to the first passage of the merger and the recent burst to the final coalescence. The 10-Myr-averaged star-formation rate inferred from the simple stellar population modelling is  $175 M_{\odot} \text{ yr}^{-1}$ , after converting to a Chabrier initial mass function<sup>62</sup>.

## Data availability

Raw data generated at the Keck Observatory are available at the Keck Observatory Archive (<https://koa.ipac.caltech.edu/>) following the standard 18-month proprietary period after the date of observation. This paper makes use of the ALMA data ADS/JAO.ALMA#2016.1.01072.S and ADS/JAO.ALMA#2017.1.01318.S, which are available at the ALMA Science Archive (<https://almascience.nrao.edu/aq/>). Some of the data presented here were obtained from the SDSS (<https://www.sdss.org>). The Hubble Space Telescope observations described here were obtained from the Hubble Legacy Archive (<https://hla.stsci.edu/>). Derived data supporting the findings of this study are available from the corresponding author upon request.

32. Tremonti, C. A., Moustakas, J. & Diamond-Stanic, A. M. The discovery of 1000 km s<sup>-1</sup> outflows in massive poststarburst galaxies at  $z = 0.6$ . *Astrophys. J.* **663**, L77–L80 (2007).
33. Rupke, D. S. N. IFSRED: data reduction for integral field spectrographs. *Astrophys. Source Code Library ascl*:**1409.004** (2014).
34. York, D. G. et al. The Sloan Digital Sky Survey: technical summary. *Astron. J.* **120**, 1579–1587 (2000).
35. Cappellari, M. & Copin, Y. Adaptive spatial binning of integral-field spectroscopic data using Voronoi tessellations. *Mon. Not. R. Astron. Soc.* **342**, 345–354 (2003).
36. Rupke, D. S. N. IFSFIT: spectral fitting for integral field spectrographs. *Astrophys. Source Code Library ascl*: **1409.005** (2014).
37. Schlafly, E. F. & Finkbeiner, D. P. Measuring reddening with Sloan Digital Sky Survey stellar spectra and recalibrating SFD. *Astrophys. J.* **737**, 103 (2011).
38. Planck Collaboration. Planck 2018 results. VI. Cosmological parameters. Preprint at <https://arxiv.org/abs/1807.06209> (2018).
39. Groves, B. A., Dopita, M. A., & Sutherland, R. S. Dusty, radiation pressure-dominated photoionization. II. Multiwavelength emission line diagnostics for narrow-line regions. *Astrophys. J.* **153** (Suppl.), 75–91 (2004).
40. Allen, M. G., Groves, B. A., Dopita, M. A., Sutherland, R. S. & Kewley, L. J. The MAPPINGS III library of fast radiative shock models. *Astrophys. J.* **178** (Suppl.), 20–55 (2008).
41. Gilli, R. et al. The X-ray to [Ne V] 3426 flux ratio: discovering heavily obscured AGN in the distant Universe. *Astron. Astrophys.* **519**, A92 (2010).
42. Vergani, D. et al. The VIMOS Public Extragalactic Redshift Survey (VIPERS). AGN feedback in [Ne V] emitters. *Astron. Astrophys.* **620**, A193 (2018).
43. Best, P. N., Röttgering, H. J. A. & Longair, M. S. Ionization, shocks and evolution of the emission-line gas of distant 3CR radio galaxies. *Mon. Not. R. Astron. Soc.* **311**, 23–36 (2000).
44. Allen, M. G., Dopita, M. A. & Tsvetanov, Z. I. Ultraviolet diagnostics for the emission-line gas in active galaxies. *Astrophys. J.* **493**, 571–582 (1998).
45. Reynaldi, V. & Feinstein, C. The fingerprints of photoionization and shock-ionization in two CSS sources. *Mon. Not. R. Astron. Soc.* **455**, 2242–2252 (2016).

46. Bolatto, A. D. et al. Suppression of star formation in the galaxy NGC 253 by a starburst-driven molecular wind. *Nature* **499**, 450–453 (2013).
47. Graham, A. W. & Driver, S. P. A concise reference to (projected) Sérsic  $R^{1/n}$  quantities, including concentration, profile slopes, Petrosian indices, and Kron magnitudes. *Publ. Astron. Soc. Aust.* **22**, 118–127 (2005).
48. Leja, J., Johnson, B. D., Conroy, C., van Dokkum, P. G. & Byler, N. Deriving physical properties from broadband photometry with Prospector: description of the model and a demonstration of its accuracy using 129 galaxies in the local Universe. *Astrophys. J.* **837**, 170 (2017).
49. Conroy, C., Gunn, J. E. & White, M. The propagation of uncertainties in stellar population synthesis modeling. I. The relevance of uncertain aspects of stellar evolution and the initial mass function to the derived physical properties of galaxies. *Astrophys. J.* **699**, 486–506 (2009).
50. Conroy, C. & Gunn, J. E. FSPS: Flexible Stellar Population Synthesis. *Astrophys. Source Code Library ascl*:**1010.043** (2010).
51. Morrissey, P. et al. The calibration and data products of GALEX. *Astrophys. J.* **173** (Suppl.), 682–697 (2007).
52. Werner, M. W. et al. The Spitzer Space Telescope Mission. *Astrophys. J.* **154** (Suppl.), 1–9 (2004).
53. Lang, D., Hogg, D. W. & Schlegel, D. J. WISE photometry for 400 million SDSS sources. *Astron. J.* **151**, 36 (2016).
54. Charlot, S. & Fall, S. M. A simple model for the absorption of starlight by dust in galaxies. *Astrophys. J.* **539**, 718–731 (2000).
55. Draine, B. T. & Li, A. Infrared emission from interstellar dust. IV. The silicate-graphite-PAH model in the post-Spitzer era. *Astrophys. J.* **657**, 810–837 (2007).
56. Stern, D. et al. Mid-infrared selection of active galactic nuclei with the Wide-Field Infrared Survey Explorer. I. Characterizing WISE-selected active galactic nuclei in COSMOS. *Astrophys. J.* **753**, 30 (2012).
57. Satyapal, S. et al. Buried AGNs in advanced mergers: mid-infrared color selection as a dual AGN candidate finder. *Astrophys. J.* **848**, 126 (2017).
58. Blecha, L., Snyder, G. F., Satyapal, S. & Ellison, S. L. The power of infrared AGN selection in mergers: a theoretical study. *Mon. Not. R. Astron. Soc.* **478**, 3056–3071 (2018).
59. Salim, S., Boquien, M. & Lee, J. C. Dust attenuation curves in the local Universe: demographics and new laws for star-forming galaxies and high-redshift analogs. *Astrophys. J.* **859**, 11 (2018).
60. Cappellari, M. & Emsellem, E. Parametric recovery of line-of-sight velocity distributions from absorption-line spectra of galaxies via penalized likelihood. *Publ. Astron. Soc. Pacif.* **116**, 138–147 (2004).
61. Cappellari, M. Improving the full spectrum fitting method: accurate convolution with Gauss-Hermite functions. *Mon. Not. R. Astron. Soc.* **466**, 798–811 (2017).
62. Chabrier, G. The Galactic disk mass function: reconciliation of the Hubble Space Telescope and nearby determinations. *Astrophys. J.* **586**, L133–L136 (2003).
63. Kauffmann, G. et al. The host galaxies of active galactic nuclei. *Mon. Not. R. Astron. Soc.* **346**, 1055–1077 (2003).
64. Kewley, L. J., Groves, B., Kauffmann, G. & Heckman, T. The host galaxies and classification of active galactic nuclei. *Mon. Not. R. Astron. Soc.* **372**, 961–976 (2006).

**Acknowledgements** We thank M. Gronke for comments on the manuscript and C. Conroy for providing the C3K models before publication. D.S.N.R. is supported in part by the J. Lester Crain Chair of Physics at Rhodes College. J.E.G. is supported by the Royal Society. This material is based upon work supported by the National Science Foundation (NSF) under a collaborative grant (AST-1814233, 1813299, 1813365, 1814159 and 1813702). We acknowledge support from NASA award number SOF-06-0191 issued by the Universities Space Research Association. Some of the data presented herein were obtained at the W. M. Keck Observatory, which is operated as a scientific partnership among the California Institute of Technology, the University of California and NASA. The Observatory was made possible by the generous financial support of the W. M. Keck Foundation. The authors wish to recognize and acknowledge the very significant cultural role and reverence that the summit of Mauna Kea has always had within the indigenous Hawaiian community. We are most fortunate to have the opportunity to conduct observations from this mountain. ALMA is a partnership of the European Southern Observatory (ESO, representing its member states), NSF (USA) and the National Institutes of Natural Sciences (Japan), together with the National Research Council (Canada), the Ministry of Science and Technology and Academia Sinica Institute of Astronomy and Astrophysics (Taiwan), and the Korea Astronomy and Space Science Institute (Republic of Korea), in cooperation with the Republic of Chile. The Joint ALMA Observatory is operated by ESO, the Associated Universities, Inc. (AUI) / National Radio Astronomy Observatory (NRAO) and the National Astronomical Observatory of Japan. NRAO is a facility of the NSF operated under cooperative agreement by AUI. The Hubble Legacy Archive is a collaboration between the Space Telescope Science Institute (STScI/NASA), the Space Telescope European Coordinating Facility (ST-ECF/ESA) and the Canadian Astronomy Data Centre (CADAC/NRC/CSA). Some of the data presented here were obtained at the MMT Observatory, a joint facility of the University of Arizona and the Smithsonian Institution. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the NSF, the US Department of Energy, NASA, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.



# Article

**Author contributions** A.C. and J.E.G. conceived the observations of a sample developed by C.T. A.C., G.L., and D.S.N.R. performed the KCWI observations, and J.E.G. led the ALMA data acquisition. D.S.N.R. led data reduction and analysis of the KCWI data, and J.E.G. led data reduction and analysis of the ALMA data. C.T. and E.R.G. fitted ancillary spectra. D.S.N.R. wrote the manuscript, with contributions from A.C. throughout; J.E.G. contributed to the section on ALMA observations; A.M.D.-S. and J.M. contributed to the section on stellar mass; and C.T. contributed to the section on stellar populations. D.S.N.R., G.L., E.R.G., J.M. and C.T. produced the figures, with A.C. and J.E.G. contributing to their design. J.M. performed the spectral energy distribution modelling, and P.H.S. handled the structural analysis of the Hubble Space

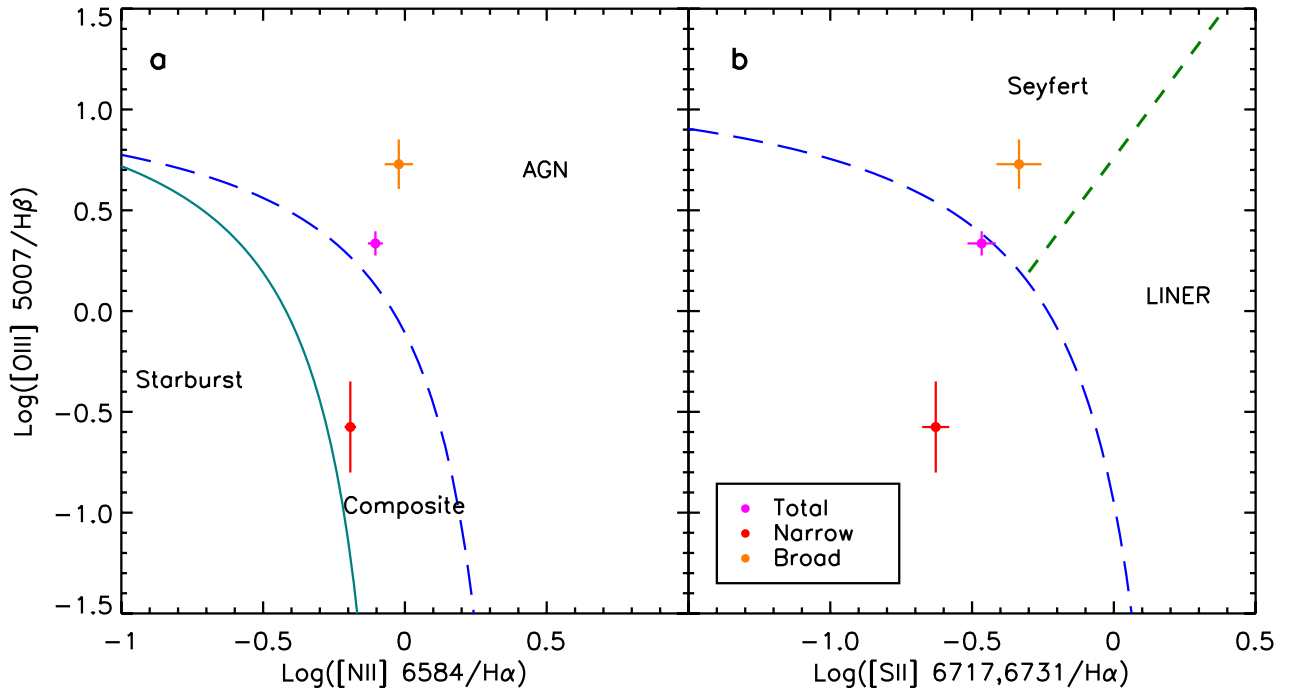
Telescope data. All co-authors provided critical feedback to the text and helped to shape the manuscript.

**Competing interests** The authors declare no competing interests.

## Additional information

**Correspondence and requests for materials** should be addressed to D.S.N.R.

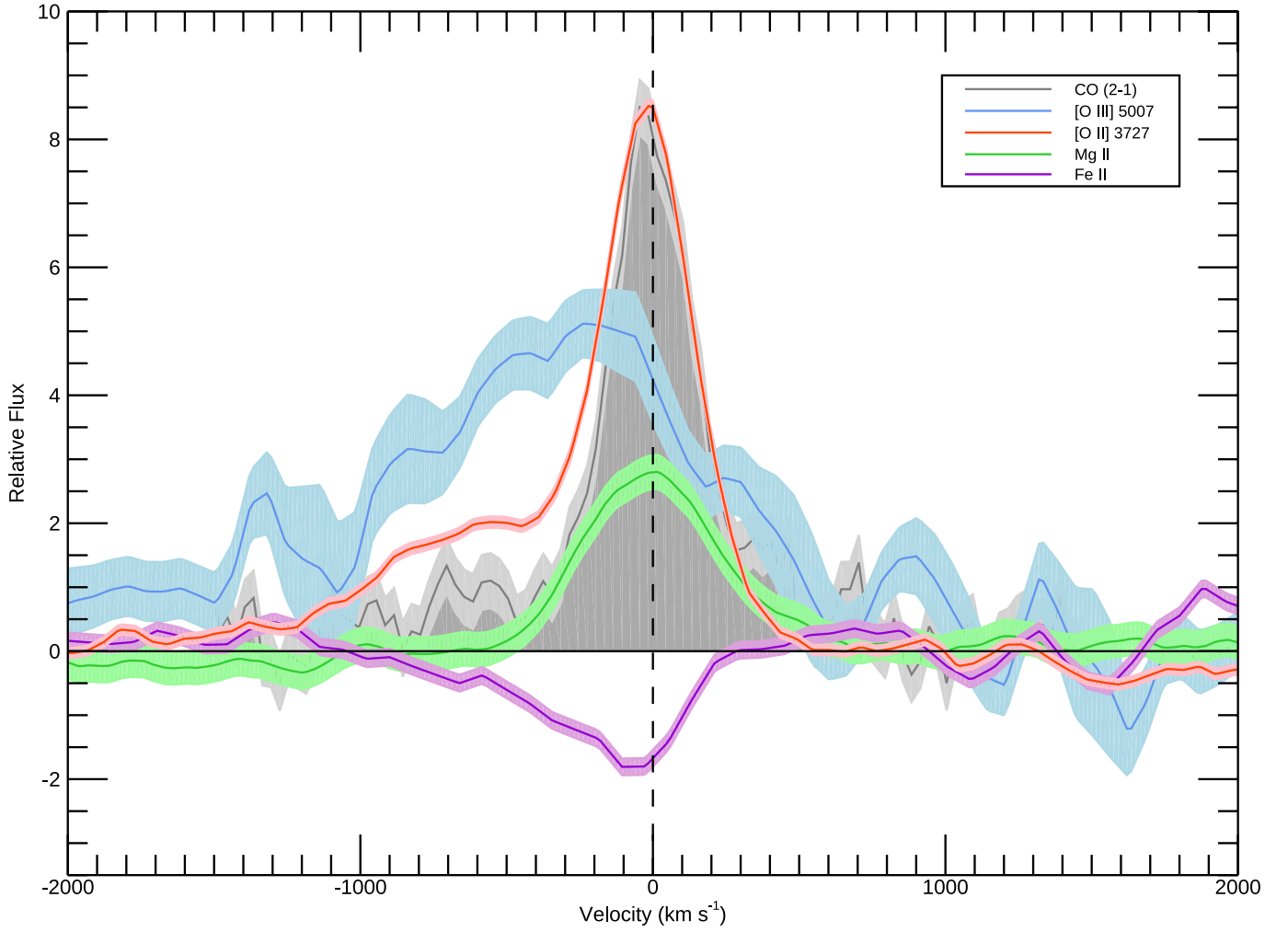
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Line ratio diagrams of the core spectrum.** In **a**, the green solid line demarcates the edge of the  $z = 0$  pure star-formation locus<sup>63</sup>; in both panels, blue long-dashed lines denote the limits of young star photoionization<sup>64</sup>; and in **b**, the green short-dashed line separates Seyfert galaxies

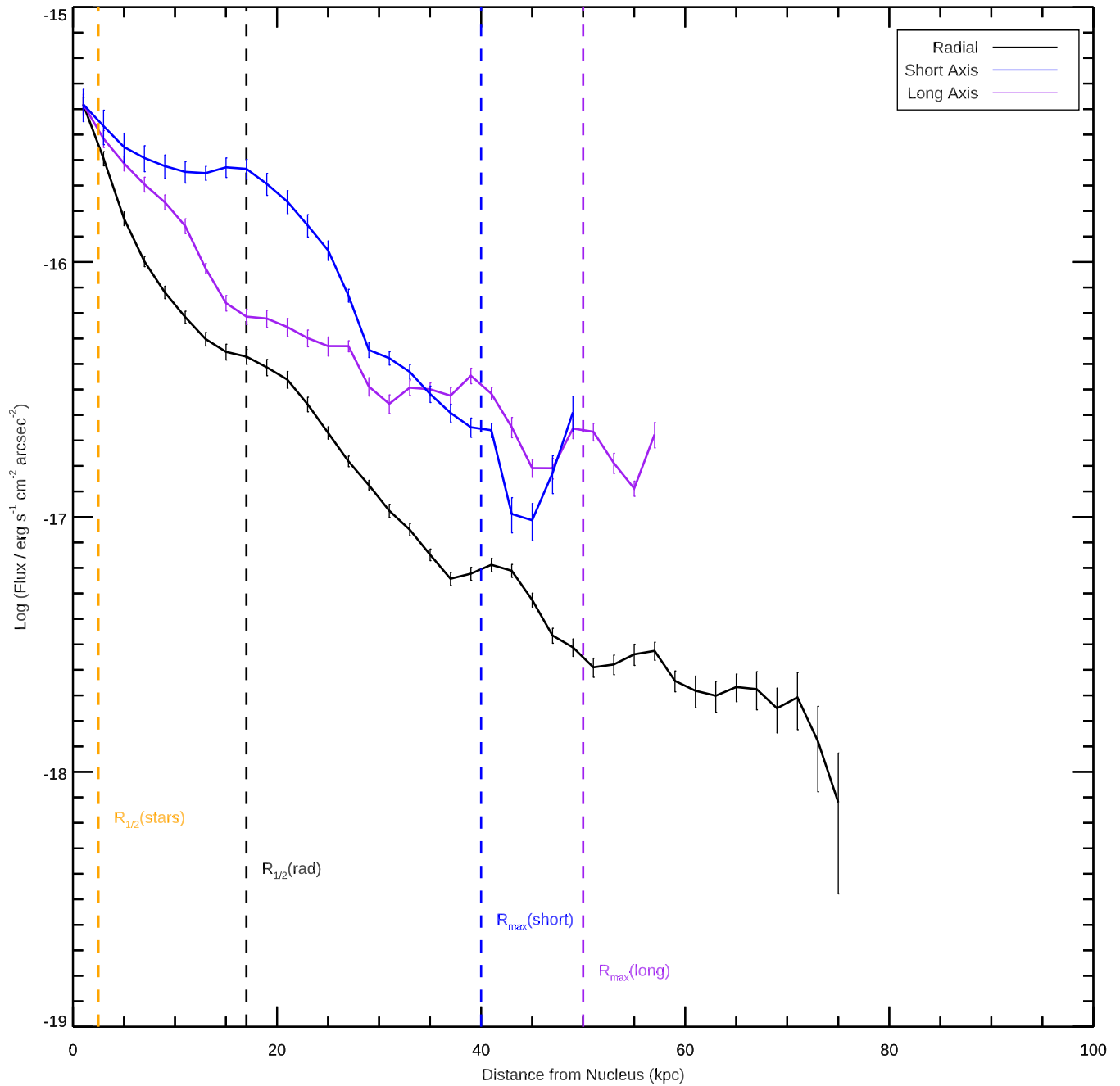
(AGNs) from low-ionization nuclear emission-line regions (LINERs)<sup>64</sup>. Error bars are  $1\sigma$ . The red narrow component is consistent with star formation at near-solar metallicity, while the broad, outflowing component is ionized by either an AGN or high-velocity shocks.





**Extended Data Fig. 2 | Comparison of velocity profiles among gas phases.** Tracers are shown as coloured lines, while the CO(2–1) profile is shaded in grey. The data are smoothed by three pixels and the coloured shadings indicate  $1\sigma$  errors on the line fluxes. The ultraviolet–optical nebular lines are shown with the

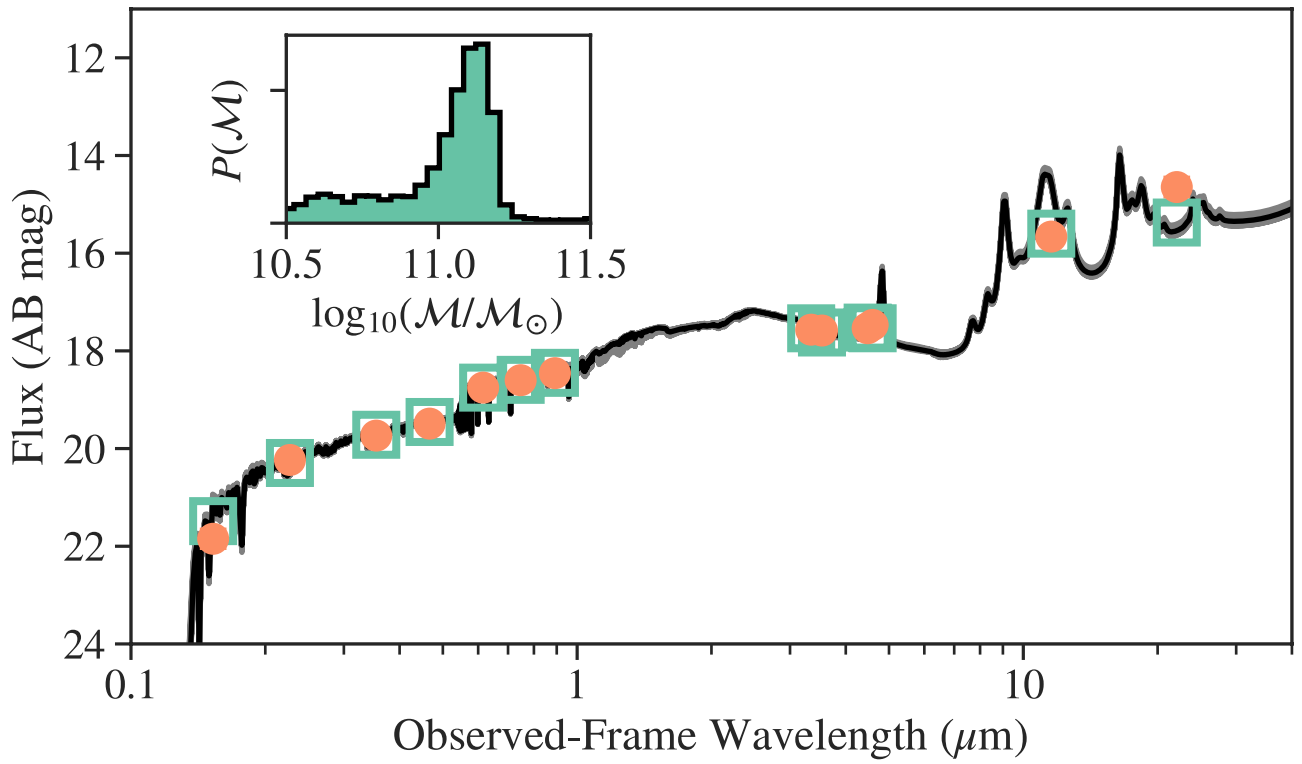
correct relative fluxes (uncorrected for reddening in the host galaxy), while the CO(2–1) line is arbitrarily scaled. The spatially integrated velocity profiles probe different gas phases and spatial scales but show remarkable overall consistency.



**Extended Data Fig. 3 | [O II] spatial profiles.** Profiles are averaged and then plotted versus distance from the galaxy nucleus along circular radii (black); the short axis of the nebula, or east-to-west axis (blue); and the long axis of the nebula, or north-to-south axis (purple). The averages are taken in directions perpendicular to these: in azimuth around the nucleus; along the long axis; and

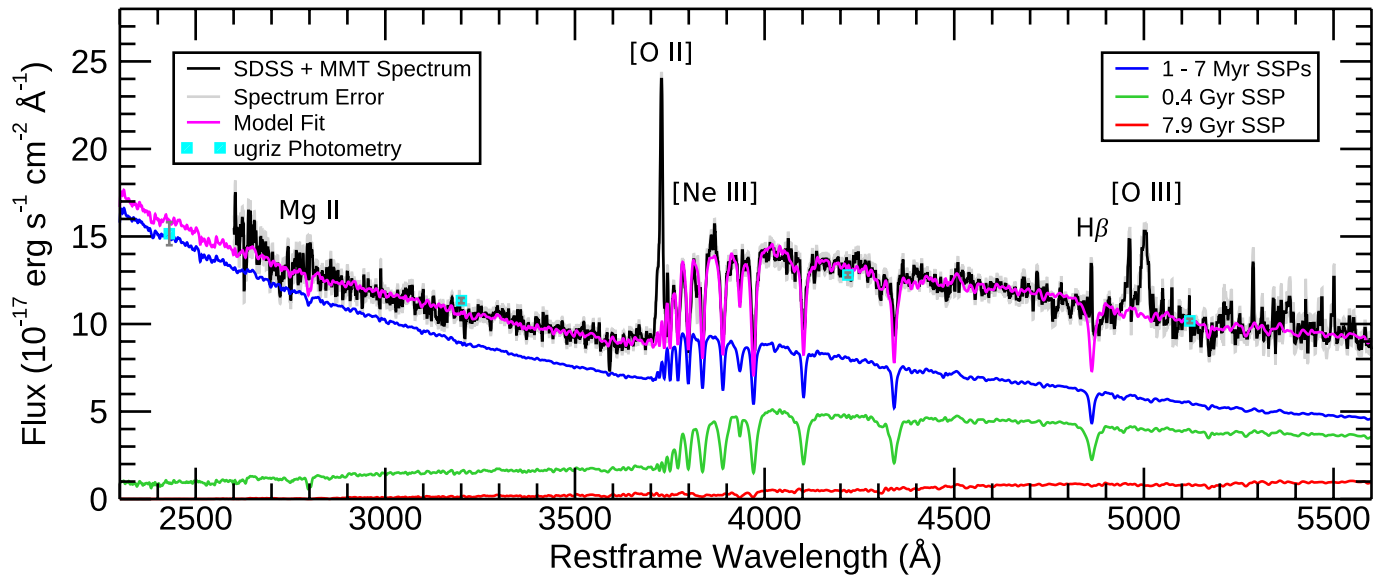
along the short axis, respectively. The short and long axis profiles are shifted upward in flux so that the three profiles match in the lowest distance bin. Errors are standard errors of the mean. Plotted as dashed lines are the stellar half-light radius (orange), the [O II] half-light radius within 50 kpc (black), and the [O II] maximum radius along the short and long axes (blue and purple).





**Extended Data Fig. 4 | Fit to the ultraviolet-to-mid-infrared spectral energy distribution.** The best-fit model and  $1\sigma$  error are shown with a black line and grey shading; observed fluxes with  $1\sigma$  errors (usually smaller than the symbols)

are yellow circles; and model fluxes are open cyan boxes. Flux is given in AB magnitudes and observed-frame wavelengths in micrometres. The posterior probability  $P(M)$  for stellar mass  $M$  is shown in the inset.



**Extended Data Fig. 5 | Stellar population model fit.** Spectral data from SDSS and the MMT and  $1\sigma$  errors are shown as the black line and grey shading. SDSS *ugriz* photometry and  $1\sigma$  errors are the cyan squares and grey vertical bars. The

best-fit model is a magenta line; the stellar population components summed to produce this model are shown as coloured lines, with ages as shown. SSP, simple stellar population.

# Braess's paradox and programmable behaviour in microfluidic networks

<https://doi.org/10.1038/s41586-019-1701-6>

Daniel J. Case<sup>1</sup>, Yifan Liu<sup>2</sup>, István Z. Kiss<sup>2</sup>, Jean-Régis Angilella<sup>3</sup> & Adilson E. Motter<sup>1,4\*</sup>

Received: 4 September 2018

Accepted: 1 August 2019

Published online: 23 October 2019

Microfluidic systems are now being designed with precision as miniaturized fluid manipulation devices that can execute increasingly complex tasks. However, their operation often requires numerous external control devices owing to the typically linear nature of microscale flows, which has hampered the development of integrated control mechanisms. Here we address this difficulty by designing microfluidic networks that exhibit a nonlinear relation between the applied pressure and the flow rate, which can be harnessed to switch the direction of internal flows solely by manipulating the input and/or output pressures. We show that these networks—implemented using rigid polymer channels carrying water—exhibit an experimentally supported fluid analogue of Braess's paradox, in which closing an intermediate channel results in a higher, rather than lower, total flow rate. The harnessed behaviour is scalable and can be used to implement flow routing with multiple switches. These findings have the potential to advance the development of built-in control mechanisms in microfluidic networks, thereby facilitating the creation of portable systems and enabling novel applications in areas ranging from wearable healthcare technologies to deployable space systems.

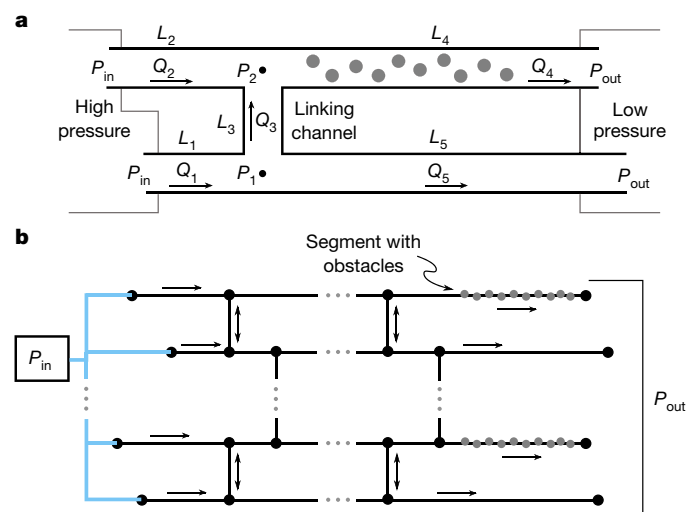
Fulfilment of the promise of microfluidics to operate as autonomous microscale networks in which fluids can be transported, mixed, reacted, separated and processed is no longer limited by experimental fabrication challenges, but rather by difficulties in creating built-in controls<sup>1–3</sup>. The importance of this limitation can be appreciated by noting that the development of the modern microelectronics that form the basis of computer microprocessors was ultimately determined by the creation of integrated circuits, with all components fabricated on the same substrate. Microfluidics have already reached a level of integration in which networks with thousands of components, including control devices, are built on a single compact chip. However, in contrast with electronic integrated circuits, existing on-chip fluid control devices still need to be actuated externally. For example, microfluidic circuits fabricated from flexible polydimethylsiloxane (PDMS) can now incorporate a large number of control valves, which nevertheless have to be operated using control fluids through a control layer that lies on top of the working fluid network<sup>4,5</sup>. As a result, microfluidics are still predominantly controlled by external hardware, despite substantial efforts over the past 20 years to develop systems with new control schemes<sup>6–10</sup>. The construction of systems that forgo the current reliance on external hardware is crucial to further the development of portable microfluidic systems for pressing applications, ranging from point-of-care diagnostics and health monitoring wearables to analysis kits for field research<sup>11–14</sup>. This requires developing next-generation integrated circuits in which not only the control devices but also the operation of those devices is integrated on-chip. The development of such a level of integration has been fundamentally limited by the fact that, at the microscale, fluid flows tend to respond linearly to pressure changes and thus cannot be easily amplified or switched.

In this Article, we explore new physics that emerges by combining network theory and fluid mechanics to induce nonlinear behaviour in microfluidics and effectively create a passive two-terminal flow-switch device that is entirely operated on-chip, directly by the working fluid. Previous work that has achieved built-in control capabilities (often externally actuated), including oscillatory flows<sup>15–18</sup> and flow rate regulation<sup>19,20</sup>, generally relied on flexible membranes and surfaces. Microfluidics with such flexible components require flows with very low Reynolds numbers—a regime in which fluid inertia, and thus the only nonlinear term of the Navier–Stokes equations for incompressible fluids, becomes negligible. This has led researchers to often discount the potential effects of fluid inertia on the flows (as reviewed, for example, in refs. <sup>21,22</sup>). Recent work has shown, however, that inertial forces can serve as a powerful on-chip tool to manipulate microfluidic dynamics locally<sup>23,24</sup>, including shaping streamlines<sup>25,26</sup>, mixing fluids<sup>27</sup> and directing particles<sup>28,29</sup>. Here, we present networks designed to amplify inertial effects by incorporating properties of porous media that can be used for non-local fluid routing and manipulation of output patterns.

Figure 1a shows a schematic representation of a microfluidic system with the fundamental network structure we consider. It consists of five segments arranged as two parallel channels connected by a linking channel, where the inlets are kept at a common pressure  $P_{\text{in}}$  and the outlets are held at a common, lower, pressure,  $P_{\text{out}}$ . One of the outlet channels is modified to generate a nonlinear pressure–flow relationship, which is achieved by introducing an array of cylindrical obstacles. Our principal results are supported by theory, simulations and experiments, and they show that we can: (i) induce a flow direction switch through the linking channel solely by varying the pressure

<sup>1</sup>Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA. <sup>2</sup>Department of Chemistry, Saint Louis University, St Louis, MO, USA. <sup>3</sup>Normandie Université, UNICAEN, UNIROUEN, ABTE, Caen, France. <sup>4</sup>Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA. \*e-mail: motter@northwestern.edu





**Fig. 1 | System schematics. a**, Microfluidic network consisting of two parallel channels, joined by a linking channel, that connect high- and low-pressure fluid reservoirs. Grey filled circles represent stationary cylindrical obstacles. The labels denote pressures ( $P$ ), channel lengths ( $L$ ) and flow rates ( $Q$ ), with arrows indicating the positive flow direction. **b**, Generic multiswitch microfluidic network consisting of an array of parallel channels interconnected by multiple linking channels. A subset of channel segments contain cylindrical obstacles. Flow is driven through the network by a single pressure difference ( $P_{in} - P_{out}$ ).

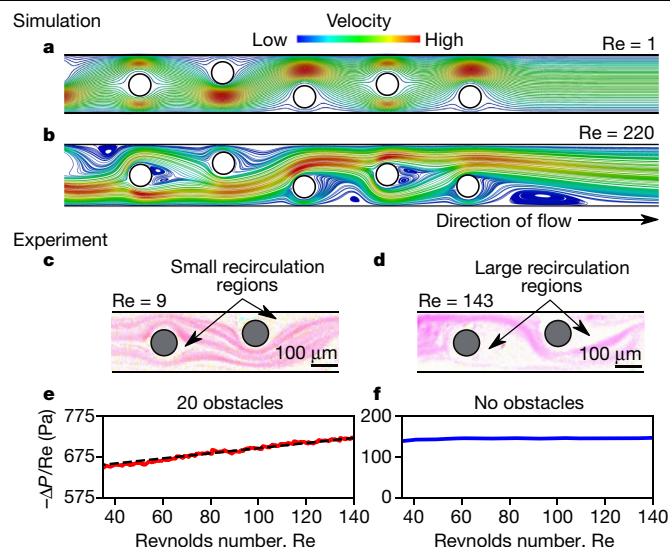
difference between the inlets and outlets; and (ii) identify a pressure difference above which the total flow rate between the inlets and outlets increases on closing the linking channel. We also predict negative conductance transitions when the linking channel is equipped with an offset fluidic diode, which are transitions associated with non-monotonic pressure–flow relations analogous to those previously realized using flexible diaphragm valves<sup>30</sup>. The counter-intuitive behaviour described in (ii) is formally equivalent to the so-called Braess’s paradox originally established for traffic networks<sup>31,32</sup>, where closing a shortcut road has the possible effect of increasing net traffic flow. We demonstrate integration of the flow switch described in (i) by considering larger microfluidic networks, as illustrated in Fig. 1b, which incorporate multiple linking channels and are thus capable of exhibiting multiple flow switches. Flows through these networks are driven by a single pressure difference and yet can be designed to exhibit a variety of flow states by programming the pressure at which each flow switch occurs.

### System design and nonlinearity

We consider conditions under which all channel segments have the same width  $w$ , the working fluid is water, and all surfaces (including obstacles) have no-slip boundaries. We assume, without loss of generality, that the pressure  $P_{out}$  at the outlets is zero, and consider scenarios in which either the static or the total pressure is controlled at the inlets (Methods). We examine two network configurations of the system in Fig. 1a: the connected configuration, in which the two parallel channels are allowed to exchange fluid through the linking channel; and the disconnected configuration, in which the linking channel is closed or removed. In our theoretical analysis and simulations, the flows are assumed to be two-dimensional, yet the main results carry over to three dimensions, as verified in our experiments.

For a straight microfluidic channel of length  $L \gg w$  without obstacles, an approximate steady-state solution of the Navier–Stokes equations in two dimensions yields a linear relation between the total volumetric flow rate per unit depth  $Q$  and the pressure drop  $\Delta P$  along the channel:

$$-\Delta P = \frac{12\mu L}{w^3} Q \quad (1)$$



**Fig. 2 | Development of nonlinear flow. a, b**, Simulated flow in a channel with obstacles (open circles), showing no recirculation for low  $Re$  (**a**) and noticeable recirculation near the obstacles for larger  $Re$  (**b**). **c, d**, Experimentally observed flows around the obstacles (filled circles), visualized using pictures of fluorescent particles (marked in pink). The particle tracks trace the underlying flow structure, confirming the development of recirculation regions (white areas) as  $Re$  is increased from low (**c**) to moderate values (**d**). **e, f**, Experimentally measured relation between pressure loss and  $Re$  for a channel with (**e**; red curve) and without (**f**; blue curve) obstacles. The dashed line in **e** is a reference to guide the eye and indicates an approximately quadratic relation between pressure loss and flow rate.

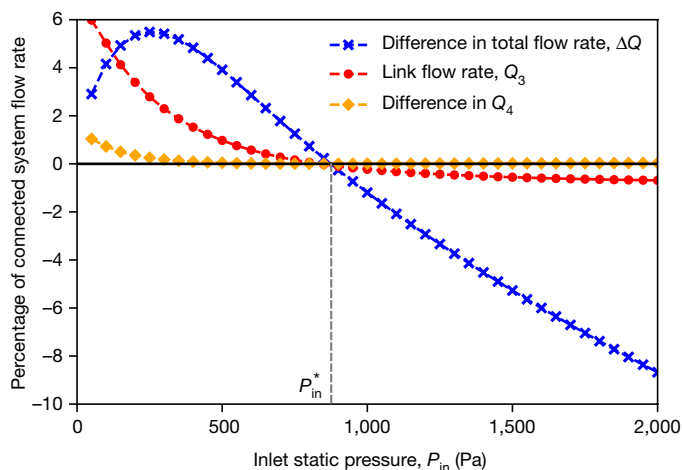
where  $\mu$  is the dynamic viscosity of the fluid. To induce deviations from this linear regime, we consider the effect of introducing multiple stationary obstacles in the channel. Figure 2a, b shows simulations of the Navier–Stokes equations for a channel with ten cylindrical obstacles of radius  $r = w/5$  (Methods). We observe recirculation regions forming near the obstacles for sufficiently large Reynolds number  $Re \equiv 2\rho Q/\mu$ , where  $\rho$  is the fluid density. The recirculation regions first appear for  $Re$  of the order of 10, and their number and size depend on  $Re$ . These localized structures are hallmarks of fluid inertia effects (and thereby of nonlinearity). We investigate how fluid inertia effects compound to impact the total flow rate by performing simulations across moderate values of  $Re$  when different numbers of obstacles are present. We find that a nonlinear relation between the pressure drop  $\Delta P$  and flow rate  $Q = \mu Re/2\rho$  emerges as soon as obstacles are introduced, and that the nonlinearity becomes more pronounced as the number of obstacles is increased (Supplementary Information section S3.1 and Supplementary Fig. 3).

The nonlinearity we observe in the relation between  $\Delta P$  and  $Q$  conforms to the Forchheimer effect in porous media, which characterizes flow through many interconnected microchannels when local inertial effects at the points of interconnection are non-negligible, even for laminar flow<sup>33–35</sup>. We use the Forchheimer equation to derive a relation between  $\Delta P$  and  $Re$  for the channel with obstacles, given by

$$-\Delta P = \frac{\alpha\mu^2 L}{2\rho w} Re + \frac{\beta\mu^2 L}{4\rho w^2} Re^2 \quad (2)$$

where  $\alpha$  is the reciprocal permeability and  $\beta$  is the non-Darcy flow coefficient, both depending solely on the system geometry (Methods).

The physical mechanism giving rise to this nonlinearity is the increase in flow recirculation and velocity gradients for larger  $Re$ , as evidenced in Fig. 2a, b for  $Re = 1$  and 220. To test the impact of the inertial effects



**Fig. 3 | Braess's paradox and flow switching.** Simulation results for the connected and disconnected configurations of the system for a range of inlet pressures  $P_{in}$ . The flow rates are presented as a percentage of the total flow rate through the connected system,  $Q_c$ , where we adopt the sign convention for the flow directions as defined in Fig. 1a. The flow through the linking channel switches direction at the critical pressure  $P_{in} = P_{in}^*$ , which coincides with the onset of negative  $\Delta Q$  that marks the occurrence of Braess's paradox.

in realistic systems, we perform experiments using microchannels fabricated from stiff PDMS (hardened by curing). Figure 2c, d shows experimental evidence of the increase in the number and size of the recirculation regions with  $Re$ , in agreement with our simulations. An approximately linear relation between  $-\Delta P/Re$  and  $Re$  (and thus an approximately quadratic relation between  $-\Delta P$  and  $Q$ ) for a channel containing 20 obstacles is shown in Fig. 2e, which contrasts with the constant relation measured for a channel without obstacles in Fig. 2f.

### Switching and Braess's paradox

We incorporate the channel segment with obstacles characterized above into a network by considering the microfluidic system presented in Fig. 1a. We take the common static pressure  $P_{in}$  at the inlets to be the controlled variable in the system. The total flow rate through the network is now simply the sum of the flows at the outlets, ( $Q_4 + Q_3$ ). In Fig. 3, we present results for this system from direct simulations of the steady-state solutions of the Navier–Stokes equations. As  $P_{in}$  is increased from zero, the flow rate through the linking channel  $Q_3$  is initially positive before changing direction and becoming negative once a critical pressure, defined as  $P_{in}^*$ , is reached (Fig. 3). This flow switch results from the nonlinear change in pressure loss along the channel segment containing obstacles, which causes a switch in the sign of the pressure difference along the linking channel  $\Delta P_{21}$  (approximately  $P_2 - P_1$ ) as the flow rate through the system increases with  $P_{in}$ . We define  $Q_c$  to be the total flow rate for the connected system configuration and  $Q_d$  to be the total flow rate for the disconnected system configuration, where both are regarded as functions of  $P_{in}$ .

Figure 3 shows  $\Delta Q \equiv Q_c - Q_d$  for a range of applied pressures  $P_{in}$ . Intuition may suggest that  $\Delta Q$  is positive for all values of  $P_{in}$  because the linking channel in the disconnected system can be considered to have an infinite fluidic resistance, while for the connected system configuration the resistance of the linking channel is finite. Hence, reducing the resistance of any component of the system may seem to imply that the total flow rate should increase for fixed  $P_{in}$ . We observe, however, that  $\Delta Q$  becomes negative for  $P_{in}$  above the critical pressure that marks the flow switch,  $P_{in}^*$ , meaning that an open linking channel between the parallel channels results in a lower total flow rate. Figure 3 also shows that the flow rate through the channel segment with obstacles,  $Q_4$ , remains largely unchanged between the two configurations. Therefore, the

difference in the total flow rate exists primarily in the difference in  $Q_3$ , and  $Q_3$  acts as a controlling variable of  $Q_5$ .

The observation of a lower total flow rate for the connected configuration compared to the disconnected configuration for fixed  $P_{in}$  is a manifestation of a fluid analogue of Braess's paradox. Indeed, if we consider the disconnected system driven by an inlet pressure  $P_{in} > P_{in}^*$ , the addition of the linking channel can result in a decrease in the total steady-state flow rate (as large as 10% in our simulations). The value of the critical pressure  $P_{in}^*$  depends, of course, on the dimensions of the channels, but we find that the onset of Braess's paradox and the flow switch always occur at the same pressure for the range of parameters investigated. We obtain similar results for Braess's paradox and flow switching when instead the total pressure is controlled at the inlets (Supplementary Information section S3.4). Our observation of Braess's paradox and flow switching also has the potential to lead to additional control features when existing microfluidic components are integrated into our system. For example, by incorporating an offset fluidic diode<sup>36</sup> in the linking channel, the system can undergo negative (and positive) conductance transitions, where an increase in  $P_{in}$  leads to an abrupt decrease in the total flow rate (Supplementary Information section S4).

### Experimental results

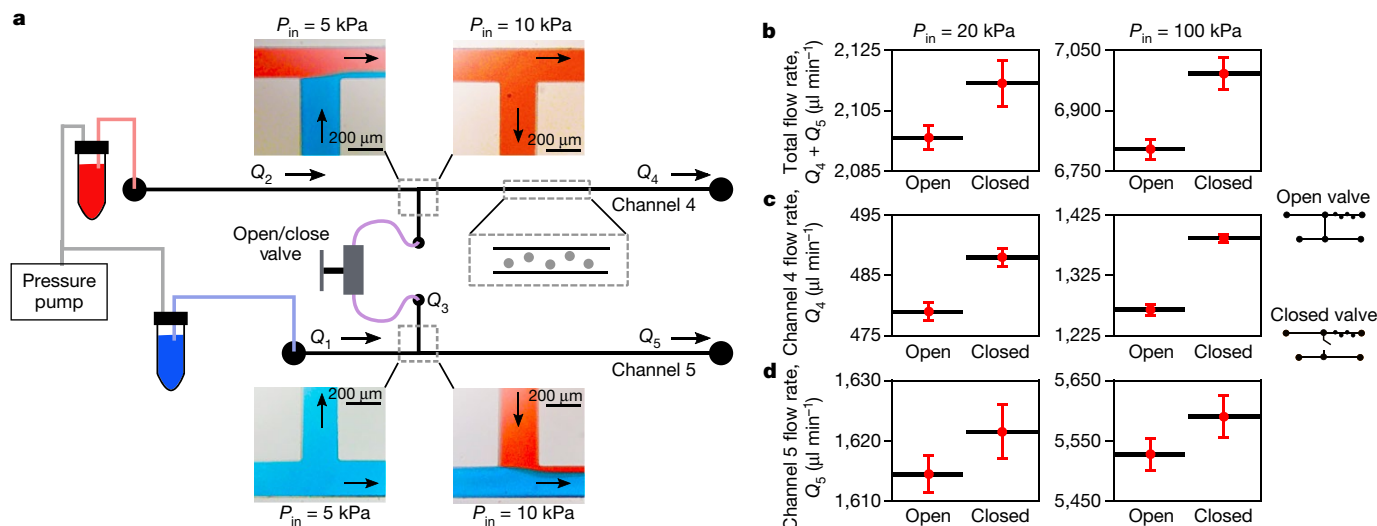
We performed experiments to validate our predictions of flow switching and Braess's paradox in a network with dimensions typical of microfluidics. A schematic of the experimental apparatus is presented in Fig. 4a, where an open/close valve is used to implement the addition/removal of the linking channel (Methods). With the valve open, a flow switch is observed at a critical driving pressure  $P_{in}^*$  in the range 5–10 kPa, as demonstrated in Fig. 4a by images of the flows through the channel junctions at the end points of this pressure range. (The switching behaviour has no reliance on the valve, as explicitly shown in Supplementary Fig. 11.)

A confirmation of Braess's paradox in this system is shown in Fig. 4b for driving pressures above  $P_{in}^*$ , as observed in our simulations. The measured total flow rate is higher when the linking channel valve is closed than when it is open, thus demonstrating the paradox, and the magnitude of the paradox is observed to be larger for higher driving pressures. A breakdown of how the flow rate changes in channel segments 4 and 5 individually is shown in Fig. 4c, d. Closing the valve causes the flow rates through both channels to increase, which is in agreement with direct simulations and is yet another striking aspect of Braess's paradox in this system; it would be, at first, intuitive to expect that  $Q_5$  would decrease when the in-flow from the linking channel is switched off. Time series of the flow rates measured as the linking channel is sequentially opened and closed further illustrate the transitions underlying the paradox (as shown in Supplementary Fig. 12).

In our experiments, the total pressure is controlled at the inlets and the experimental results are in full qualitative agreement with simulations performed under the same pressure boundary conditions (Supplementary Information section S3.4). This illustrates the robustness of the phenomenon, given that our simulations are in two dimensions and three-dimensional effects are expected to be present in the experiments. We note that different aspects of the paradox have been considered in fluid networks, but only for macroscopic (that is, non-microfluidic) systems and while modelled by ad hoc flow equations<sup>37–39</sup>. Analogues of the paradox have also been studied in several other areas, including electrical, mechanical, biological, and contemporary traffic networks<sup>40–44</sup>. These examples show that Braess's paradox is a potentially general network phenomenon, which has remained unexplored in microfluidic networks.

### Network model

To characterize the microfluidic system in Fig. 1a, we construct an analytic model that captures the flow properties observed in our simulations and experiments. The model consists of pressure–flow relations



**Fig. 4 | Experimental observation of flow switch and Braess's paradox.**

**a**, Experimental setup of the system presented in Fig. 1a, with flow tracking images (insets) at the junctions. An air-pressure pump is used to equally pressurize two vials containing red and blue dyed water, where each vial is connected to one of the system inlets. The linking channel is equipped with an open/close valve and channel 4 contains 20 obstacles. Images of the dyed flows through the junctions are shown for  $P_{in}$  below (5 kPa; left insets) and above (10 kPa;

right insets) the flow switching pressure  $P_{in}^*$ , where the flow directions are indicated by the arrows. **b**, Total flow rate ( $Q_4 + Q_5$ ) when the linking channel valve is 'open' or 'closed' (see diagrams at right) for two different driving pressures above  $P_{in}^*$ . **c**, **d**, Breakdown of the total flow rate into  $Q_4$  (**c**) and  $Q_5$  (**d**) for the two states of the valve. The plotted flow rates are averages derived from time series data, and the error bars indicate one standard deviation. The observed increase in the total flow rate when the valve is closed is direct evidence of Braess's paradox.

for each channel segment and, crucially, includes the most dominant term resulting from minor pressure losses at the channel junctions<sup>45,46</sup> (Methods). We model the contribution of the latter as an additive term  $K(Q_3/Q_1)f(Q_3)$  to the pressure–flow equation for channel segment 5, where the scaling factor  $f$  and the coefficient  $K$  are increasing functions for  $P_{in} \geq 0$  such that  $f(0) = K(0) = 0$ . Several results are obtained from this model for  $P_{in} > 0$ , as assumed throughout. First, if  $\beta = 0$  (that is, the quadratic term is zero in equation (2)) when the static pressure is controlled or the dynamic pressure is negligible, then flow switching does not occur, in agreement with direct simulations (Supplementary Information section S3.2). Second, when  $\beta > 0$ , a steady-state solution can be found satisfying  $Q_3 = 0$  provided that the following geometric condition is satisfied:

$$L_1 < \frac{12L_2L_5}{\alpha\omega^2L_4} = L^* \quad (3)$$

This solution identifies the critical pressure  $P_{in}^*$ . Third, for flow rates in the linking channel, the model predicts that a variation  $\delta Q_3$  is negatively related to a variation  $\delta P_{in}$  around  $P_{in}^*$ . This indicates that  $P_{in}$  above (below)  $P_{in}^*$  results in a negative (positive) flow rate through the linking channel. The first result implies that, in our experiments, the Forchheimer effect is necessary to achieve a flow switch. The second and third results, which hold even for when dynamic pressure is non-negligible, show that this model captures the flow switching behaviour observed in the simulations and experiments. Importantly, we validate the flow-switching condition in equation (3) by demonstrating quantitative agreement between the model and simulations both when the static pressure and when the total pressure is controlled (Supplementary Information section S3.2).

The model also predicts Braess's paradox as observed in our experiments and simulations. Specifically, under the condition that equation (3) is satisfied and dynamic pressure is small (or static pressure is controlled), the model predicts the paradox to occur for  $\delta P_{in} > 0$  if and only if

$$K'(0)\beta f\left(\frac{a}{\beta}\right) > c \quad (4)$$

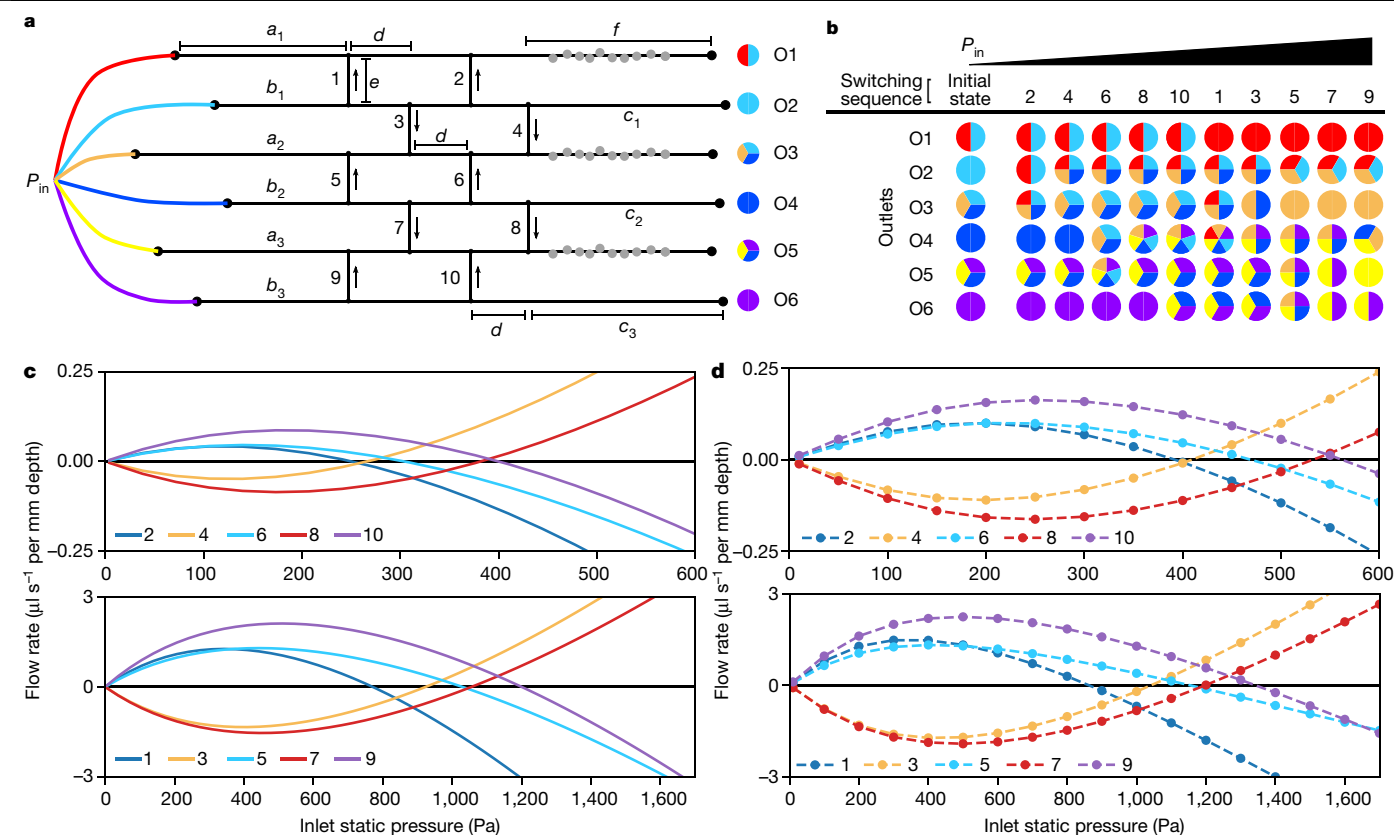
where  $a$  and  $c$  are positive parameters and prime denotes derivative. If total pressure is controlled and dynamic pressure terms are included, the paradox is also predicted for  $\delta P_{in} > 0$  provided that a relation similar to equation (4) is satisfied (details for both cases are presented in Supplementary Information section S2). The dependence of condition (4) on  $\beta$  and  $K'(0)$  underlines the crucial roles of nonlinearity and minor losses in giving rise to Braess's paradox in our experiments, and shows in particular that minor losses have to be sufficiently large. Indeed, if the effect of minor losses is neglected, a manifestation of Braess's paradox is still predicted to occur, but with much smaller magnitude and only for  $\delta P_{in} < 0$ , which is inconsistent with our simulations and experiments (Supplementary Information section S2.3).

The result in equation (4) also highlights a fundamental difference between microfluidic and electronic circuits, namely that minor losses (that is, energy losses associated with interactions between circuit components) do not have direct analogues in common electronics. Given the central role played by such losses in equation (4), we posit that this difference might be the reason why no equivalent of the Braess paradox effect we present has been observed in electrical networks, even though aspects of it have<sup>40</sup>. We further investigated the impact of interactions between channel segments by varying the junction angles to show that the paradox can be further enhanced by manipulating the minor losses (Supplementary Information section S3.3).

### Networks with multiple programmed switches

The system considered thus far can be generalized to create larger microfluidic networks with multiple flow switches—that is, networks with multiple disjoint channel segments in which the flow initially in one direction can be individually 'switched' to move in the opposite direction through the manipulation of one driving pressure alone. In our design, the linking channel plays the role of a switch (and can be referred to as such). Figure 1b shows the multiswitch generalization of the network in Fig. 1a, which incorporates multiple linking channels and a subset of channel segments with obstacles. We experimentally demonstrate an instance of a six-switch network that exhibits flow switching in all linking channels (as presented in Supplementary Information section





**Fig. 5 | Flow patterns in a multiswitch network.** **a**, Schematic of ten-switch network. Fluids of different colours are driven to each inlet by a common static pressure source,  $P_{in}$ . The outlets are labelled by O1–O6 and the linking channels by 1–10. The arrows indicate the flow direction through each linking channel and multicoloured circles schematically indicate the fluid composition at each outlet for an initially low  $P_{in}$ . The segment lengths are denoted by  $a_i, b_i, c_i, d, e$  and  $f$ , where a common length is assumed for all linking channels and the segments with obstacles are marked with filled grey circles. **b**, Patterns of outlet flows for the network programmed with a chosen switching sequence as  $P_{in}$  is increased. Each column of coloured circles denotes the outlet flows after the corresponding flow switch occurs, where mixing between different coloured

fluids is assumed to occur when passing through the same channel segment. **c, d**, Model predictions (c) and simulation results of the Navier–Stokes equations (d) for the flow rate through each linking channel for a network designed to exhibit the switching sequence in b. The flow rates are labelled according to the channels in a and are divided into two sets (top and bottom panels) for clarity. Positive flow rates correspond to flow in the upward direction in a, and each flow switch occurs when the corresponding curve crosses the horizontal axis. The segment dimensions that give rise to the particular switching order in b–d are reported in Supplementary Table 1. All 21 possible outlet flow colour combinations are realized between the switching sequence presented here and those in Supplementary Fig. 13.

S6.2). Multiswitch networks can be designed by extending the network model presented above.

One such network with ten linking channels is presented in Fig. 5a. By marking each inlet flow with a different colour, we show that a variety of patterns can form in the outlet flows (coloured circles in Fig. 5). The specific pattern at an outlet depends on the order in which the flow switches occur as  $P_{in}$  is varied. The network model for larger systems is constructed by combining pressure–flow relations for each channel segment with flow rate conservation equations for each junction. Using this model, we can design a network for which each flow switch occurs near a target value of  $P_{in}$  by optimizing the dimensions of the channel segments (Methods).

As illustrated in Fig. 5, a set of 11 different internal flow states and 17 unique colour combinations at the outlets are possible for the switching sequence realized in Fig. 5b. Figure 5c, d shows the agreement between the model predictions of these flow states and results from direct simulations of the Navier–Stokes equations. This variety of states (and output patterns) is achieved with only three channel segments containing obstacles and is parameterized by a single control variable—the driving pressure  $P_{in}$ . Moreover, the switching is implemented solely through the working fluid, which differs from existing approaches that rely on flexible valves and additional control flows<sup>15</sup>. Thus, multiswitch networks exhibit several properties exploitable in the design of new controllable microfluidic systems.

More generally, for a multiswitch network with  $n_c$  horizontal channels interconnected by  $n_l$  linking channels, the number of possible internal flow states is  $n_l + 1$  if each linking channel exhibits a flow switch. In addition, the possible number of unique colour combinations in the outlet flows is  $n_c(n_c + 1)/2$  if each inlet flow is marked with a different colour. All colour combinations can be realized over the set of all switching sequences, provided that there exist flow paths allowing mixing of every set of  $k$  adjacent colours for  $k$  ranging from 1 to  $n_c$ . The myriad states possible in such multiswitch networks underlie their ability to process inputs into multiple outputs and thus to support various applications, including implementing different mixing orders of chemical reagents and devising schemes for the parallel generation of mixtures with tunable concentrations.

## Conclusions and outlook

The flow switch, conductance transitions and Braess’s paradox established in this study are all emergent behaviours of common origin resulting from nonlinearity and interactions between different parts of the system. The nonlinearity is directly determined by fluid inertia effects, which can be enhanced and manipulated through the placement of obstacles and has the advantage of not being reliant on flexible components, fluid compressibility or dedicated control flows. The onset of Braess’s paradox is marked by the flow-switching pressure, above which the increased resistance of the nonlinear channel causes the flow to be

routed in the negative direction through the linking channel. When constrained by a diode, the switch in flow direction also enables negative conductance transitions. Our results demonstrate an approach for routing and switching in microfluidic networks through control mechanisms that are coded into the network structure, thus responding to the call for design strategies that allow diverse microfluidic systems to be assembled from a small set of core components<sup>2,47</sup>.

Here we considered the scenario in which the inlets and the outlets are (separately) held at the same pressure, rendering the network a two-terminal system in all cases, because this is the most stringent scenario for flow manipulation. If a multi-terminal system is configured, by allowing the pressures at each of the inlets (and/or outlets) to be varied independently, then the effects that we presented may be further enhanced. Finally, although we focused on boundary conditions in which the inlet pressures are controlled, it would be natural to explore in future research the scenario in which the controlled variables are the inlet flow rates. We anticipate, for example, that the negative conductance transitions would then be converted into pressure amplification (pressure release) transitions in which the inlet–outlet pressure difference increases (decreases) abruptly at the transition point. Accordingly, Braess's paradox is also expected to take a complementary form in which closing the linking channel causes the inlet–outlet pressure difference to drop. Incidentally, it is this complementary form of Braess's paradox that has been previously established for electrical circuits<sup>40</sup>, thus suggesting an additional correspondence between electronic and microfluidic circuits.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1701-6>.

- Pennathur, S. Flow control in microfluidics: are the workhorse flows adequate? *Lab Chip* **8**, 383–387 (2008).
- Stone, H. A. Microfluidics: tuned-in flow control. *Nat. Phys.* **5**, 178–179 (2009).
- Perdigones, F., Luque, A. & Quero, J. M. Correspondence between electronics and fluids in MEMS: designing microfluidic systems using electronics. *IEEE Ind. Electron. Mag.* **8**, 6–17 (2014).
- Thorsen, T., Maerkl, S. J. & Quake, S. R. Microfluidic large-scale integration. *Science* **298**, 580–584 (2002).
- Geertz, M., Shore, D. & Maerkl, S. J. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proc. Natl Acad. Sci. USA* **109**, 16540–16545 (2012).
- Seker, E. et al. Nonlinear pressure-flow relationships for passive microfluidic valves. *Lab Chip* **9**, 2691–2697 (2009).
- Weaver, J. A., Melin, J., Stark, D., Quake, S. R. & Horowitz, M. A. Static control logic for microfluidic devices using pressure-gain valves. *Nat. Phys.* **6**, 218–223 (2010).
- Tanyeri, M., Ranka, M., Sittipolkul, N. & Schroeder, C. M. Microfluidic Wheatstone bridge for rapid sample analysis. *Lab Chip* **11**, 4181–4186 (2011).
- Kim, S.-J., Lai, D., Park, J. Y., Yokokawa, R. & Takayama, S. Microfluidic automation using elastomeric valves and droplets: reducing reliance on external controllers. *Small* **8**, 2925–2934 (2012).
- Li, L., Mo, J. & Li, Z. Nanofluidic diode for simple fluids without moving parts. *Phys. Rev. Lett.* **115**, 134503 (2015).
- Chin, C. D., Linder, V. & Sia, S. K. Commercialization of microfluidic point-of-care diagnostic devices. *Lab Chip* **12**, 2118–2134 (2012).
- Araci, I. E., Su, B., Quake, S. R. & Mandel, Y. An implantable microfluidic device for self-monitoring of intraocular pressure. *Nat. Med.* **20**, 1074–1078 (2014).
- Bhatia, S. N. & Ingber, D. E. Microfluidic organs-on-chips. *Nat. Biotechnol.* **32**, 760–772 (2014).

- Sackmann, E. K., Fulton, A. L. & Beebe, D. J. The present and future role of microfluidics in biomedical research. *Nature* **507**, 181–189 (2014).
- Leslie, D. C. et al. Frequency-specific flow control in microfluidic circuits with passive elastomeric features. *Nat. Phys.* **5**, 231–235 (2009).
- Mosadegh, B. et al. Integrated elastomeric components for autonomous regulation of sequential and oscillatory flow switching in microfluidic devices. *Nat. Phys.* **6**, 433–437 (2010).
- Duncan, P. N., Nguyen, T. V. & Hui, E. E. Pneumatic oscillator circuits for timing and control of integrated microfluidics. *Proc. Natl Acad. Sci. USA* **110**, 18104–18109 (2013).
- Duncan, P. N., Ahrar, S. & Hui, E. E. Scaling of pneumatic digital logic circuits. *Lab Chip* **15**, 1360–1365 (2015).
- Doh, I. & Cho, Y.-H. Passive flow-rate regulators using pressure-dependent autonomous deflection of parallel membrane valves. *Lab Chip* **9**, 2070–2075 (2009).
- Collino, R. R. et al. Flow switching in microfluidic networks using passive features and frequency tuning. *Lab Chip* **13**, 3668–3674 (2013).
- Stroock, A. D. et al. Chaotic mixer for microchannels. *Science* **295**, 647–651 (2002).
- Squires, T. M. & Quake, S. R. Microfluidics: fluid physics at the nanoliter scale. *Rev. Mod. Phys.* **77**, 977–1026 (2005).
- Amini, H., Lee, W. & Di Carlo, D. Inertial microfluidic physics. *Lab Chip* **14**, 2739–2761 (2014).
- Zhang, J. et al. Fundamentals and applications of inertial microfluidics: a review. *Lab Chip* **16**, 10–34 (2016).
- Tesaf, V. & Bandalusena, H. C. H. Bistable diverter valve in microfluidics. *Exp. Fluids* **50**, 1225–1233 (2011).
- Amini, H. et al. Engineering fluid flow using sequenced microstructures. *Nat. Commun.* **4**, 1826 (2013).
- Sudarsan, A. P. & Ugaz, V. M. Multivortex micromixing. *Proc. Natl Acad. Sci. USA* **103**, 7228–7233 (2006).
- Di Carlo, D., Edd, J. F., Humphry, K. J., Stone, H. A. & Toner, M. Particle segregation and dynamics in confined flows. *Phys. Rev. Lett.* **102**, 094503 (2009).
- Wang, X. & Papautsky, I. Size-based microfluidic multimodal microparticle sorter. *Lab Chip* **15**, 1350–1359 (2015).
- Xia, H. M. et al. Analyzing the transition pressure and viscosity limit of a hydroelastic microfluidic oscillator. *Appl. Phys. Lett.* **104**, 024101 (2014).
- Braess, D. Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* **12**, 258–268 (1968).
- Braess, D., Nagurney, A. & Wakolbinger, T. On a paradox of traffic planning. *Transport. Sci.* **39**, 446–450 (2005).
- Rojas, S. & Koplik, J. Nonlinear flow in porous media. *Phys. Rev. E* **58**, 4776–4782 (1998).
- Andrade, J. S. Jr, Costa, U. M. S., Almeida, M. P., Makse, H. A. & Stanley, H. E. Inertial effects on fluid flow through disordered porous media. *Phys. Rev. Lett.* **82**, 5249–5252 (1999).
- Fourar, M., Radilla, G., Lenormand, R. & Moyne, C. On the non-linear behavior of a laminar single-phase flow through two and three-dimensional porous media. *Adv. Water Resour.* **27**, 669–677 (2004).
- Adams, M. L., Johnston, M. L., Scherer, A. & Quake, S. R. Polydimethylsiloxane based microfluidic diode. *J. Micromech. Microeng.* **15**, 1517–1521 (2005).
- Calvert, B. & Keady, G. Braess's paradox and power-law nonlinearities in networks. *J. Aust. Math. Soc. Ser. B* **35**, 1–22 (1993).
- Penchina, C. M. Braess's paradox and power-law nonlinearities in five-arc and six-arc two-terminal networks. *Open Transplant. J.* **3**, 8–14 (2009).
- Ayala, L. F. & Blumsack, S. The Braess paradox and its impact on natural-gas-network performance. *Oil Gas Facilities* **2**, 52–64 (2013).
- Cohen, J. E. & Horowitz, P. Paradoxical behavior of mechanical and electrical networks. *Nature* **352**, 699–701 (1991).
- Youn, H., Gastner, M. T. & Jeong, H. Price of anarchy in transportation networks: efficiency and optimality control. *Phys. Rev. Lett.* **101**, 128701 (2008).
- Nicolaou, Z. G. & Motter, A. E. Mechanical metamaterials with negative compressibility transitions. *Nat. Mater.* **11**, 608–613 (2012).
- Pala, M. G. et al. Transport inefficiency in branched-out mesoscopic networks: an analog of the Braess paradox. *Phys. Rev. Lett.* **108**, 076802 (2012).
- Motter, A. E. & Timme, M. Antagonistic phenomena in network dynamics. *Annu. Rev. Condens. Matter Phys.* **9**, 463–484 (2018).
- Crane Co. Engineering Division. *Flow of Fluids through Valves, Fittings, and Pipe*. Technical paper no. 410 (Crane Co., 2010).
- Khodaparast, S., Borhani, N. & Thome, J. R. Sudden expansions in circular microchannels: flow dynamics and pressure drop. *Microfluid. Nanofluidics* **17**, 561–572 (2014).
- Bhargava, K. C., Thompson, B. & Malmstadt, N. Discrete elements for 3D microfluidics. *Proc. Natl Acad. Sci. USA* **111**, 15013–15018 (2014).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### Navier–Stokes simulations

The numerical simulations were performed using<sup>48</sup> OpenFOAM version 4.1. We used meshes with an average cell area ranging from  $10\ \mu\text{m}^2$  to  $340\ \mu\text{m}^2$ , where the finest meshing was applied near the obstacles. All meshes were generated using Gmsh<sup>49</sup>. The two-dimensional solutions were found using the simpleFoam solver within OpenFOAM, employing second-order numerical schemes, where a fixed static pressure of zero was set for the boundary conditions at the outlets. At the inlets, the static (total) pressure was fixed for the static (total) pressure controlled cases. For simulations of the multiswitch network in Fig. 5, the same geometry and dimensions were used as for the model predictions, provided in Supplementary Table 1, and equal driving pressures were applied at each of the six inlets.

### Reynolds numbers

The characteristic length scale used in defining the Reynolds number of the flow is the hydraulic diameter of the channels, defined as  $4A/P$ , where  $A$  is the area and  $P$  is the perimeter of the channel cross-section (common to all segments). The hydraulic diameter in two and three dimensions is  $2w$  and  $2wh/(w+h)$ , respectively, where  $h$  is the height of the channels in the three-dimensional case. The characteristic velocity used in two and three dimensions is  $Q/w$  and  $Q/wh$ , respectively. Therefore, we define  $\text{Re} = 2\rho Q/\mu$  for our simulations in two dimensions and  $\text{Re} = 2\rho Q/\mu(w+h)$  for our experiments in three dimensions. The undeclared ranges of  $\text{Re}$  for the channel segment with obstacles considered in the presented data are: 21–385 (Fig. 3), 12–121 (Fig. 4), 1–220 (Fig. 5), 1–380 (Supplementary Fig. 2), 4–111 (Supplementary Fig. 4), 40–385 (Supplementary Fig. 7), 20–400 (Supplementary Fig. 8), 2–10 (Supplementary Fig. 11b), 75–85 (Supplementary Fig. 11c), 76–89 (Supplementary Fig. 12), 10–20 (Supplementary Fig. 14b) and 110–120 (Supplementary Fig. 14c).

### Pressure boundary conditions

We consider two different boundary conditions for the driving pressure  $P_{\text{in}}$  at the system inlets. Under one condition, total pressure is controlled and the inlets open directly into a high-pressure reservoir. Under the other condition, static pressure is controlled and the inlets are connected to the reservoir by pressure regulators. Total pressure is the sum of static pressure and dynamic pressure, where dynamic pressure is defined as  $\frac{1}{2}\rho v^2$  for a fluid with density  $\rho$  and velocity  $v$ . The distinction between these boundary conditions is often neglected in the microfluidics literature when the Reynolds number is less than one<sup>50</sup>, but it can become important for larger Reynolds numbers (even though the flow remains laminar)<sup>51</sup>.

### Pressure–flow relations

We use equation (1) to describe the pressure–flow relation for straight, obstacle-free channels; it is derived directly from the Navier–Stokes equations by assuming plane Poiseuille flow through a two-dimensional channel. To describe the nonlinear pressure–flow relation observed for the channel with obstacles, we refer to the Forchheimer equation:  $-\Delta P = \alpha\mu LV + \beta\rho LV^2$ , where  $V$  is the average fluid velocity. In two dimensions,  $V = Q/w = \mu\text{Re}/2\rho w$  and, thus, the Forchheimer equation can be written in the form of equation (2). In agreement with equation (2), we find an excellent linear fit between  $-\Delta P/\text{Re}$  and  $\text{Re}$  for a channel with ten obstacles, and we validate the fit by predicting flows through the same channel for a fluid with a different viscosity (Supplementary Information section S3.1). We observe no unsteady flow through the channel with obstacles due to vortex shedding for  $\text{Re}$  of up to 400, as expected for systems with highly confined obstacles<sup>52</sup>, which permits the use of the steady-state relation in equation (2) over the range of  $\text{Re}$  considered here. We experimentally verify the source of nonlinearity in PDMS channels with obstacles, which were designed to have approximately square cross-sections to minimize deformation (which could lead to other forms of

nonlinearity<sup>53,54</sup>). Through additional experiments, we confirmed that pressure–flow relations similar to those in Fig. 2e, f hold for channels constructed from materials with both higher rigidity (SU-8 photoresist) and lower rigidity (Flexdym) than the PDMS (Supplementary Information section S5 and Supplementary Fig. 10). We note that porous-like structures have been previously used both to study non-inertial effects in microfluidics, such as droplet formation<sup>55</sup> and viscous fingering<sup>56</sup>, and to study inertial effects in larger systems<sup>57</sup>. In our system, inertial effects arise at the microfluidic scale even for a much smaller number of obstacles than the typical number in porous-like materials.

### Network flow model construction

The analytic model used to describe the system in Fig. 1a is constructed as follows: (i) we consider the pressure at the inlets  $P_{\text{in}}$  to be in the vicinity of  $P_{\text{in}}^*$ ; (ii) we approximate the pressure–flow relation through the linking channel as  $Q_3 = \kappa(\gamma P_1 - P_2)$ , where  $\kappa$  is the channel conductivity and  $\gamma$  is a free parameter allowing for an effective pressure difference; (iii) the flow equation for each other channel segment without obstacles is written as in equation (1), where  $-\Delta P$  is the pressure drop along the segment and  $L$  is the segment length; (iv) for the channel segment with obstacles, we take the flow equation to be in the form of equation (2) (with  $\text{Re}$  expressed as  $2\rho Q/\mu$ ); (v) we include the most dominant term resulting from minor pressure losses at the channel junctions. Therefore, the model consists of five pressure–flow relations, in addition to two flow conservation equations at the junctions:  $Q_3 + Q_2 - Q_4 = 0$  and  $Q_3 + Q_5 - Q_1 = 0$ . When the static pressure is controlled at the inlets, the only nonlinearity that exists in the model comes from the Forchheimer term due to the presence of obstacles and the minor loss term. The model can also be adapted for when total pressure is controlled by taking the static pressure at each inlet to be  $P_{\text{in}} - k\rho Q^2/2w^2$ , where  $P_{\text{in}}$  now denotes total pressure and the coefficient  $k$  is a constant of order unity that only depends on the shape of the inlet velocity profile ( $k \approx 1$  for a uniform velocity profile at the inlet, as considered here). However, the dynamic pressure term  $\rho Q^2/2w^2$  is often negligible in real microfluidic systems because of the high pressures needed to drive fluid through the channels. Indeed, in our experiments, the dynamic pressure near  $P_{\text{in}}^*$  was smaller than the static pressure by two orders of magnitude and smaller than the pressure loss due to the Forchheimer effect by one order of magnitude. This can also be seen in Fig. 2f, where a constant relation between  $\text{Re}$  and  $\Delta P/\text{Re}$  is measured. Details of the model are presented in Supplementary Information section S1.

### Designing multiswitch networks

For a network with multiple switches and a given set of channel dimensions, the value of  $P_{\text{in}}$  for which a specific flow switch occurs can be determined through the addition of a constraint to the model that enforces the flow through the corresponding linking channel to be zero. Then, the dimensions of a chosen subset of channel segments may be varied through an optimization procedure in order to design a network for which each flow switch occurs near a target value of  $P_{\text{in}}$ . Depending on which dimensions are allowed to be adjusted, the desired relative order of the switches can be achieved exactly, and the final set of switching pressures can be very close to the target ones (often <5% difference), where the former is expected to be more important in applications. Further details on the design of multiswitch networks are presented in Supplementary Information section S6.1.

### PDMS channel fabrication

The flow channels were assembled by sealing a patterned PDMS chip against a glass slide. The PDMS chip was made by pouring a mixture of PDMS oligomer and cross-linking curing agent (Sylgard 184) at a weight ratio of 10:1 into a mould after being degassed under vacuum. The mixture was cured at 74 °C for 1 h and then peeled off from the mould to yield the microchannel design. The dimensions of the channels in Figs. 2 and 4 were  $200\ \mu\text{m}$  (width)  $\times$   $185\ \mu\text{m}$  (height), and the diameter of the obstacles was



97  $\mu\text{m}$ . After punching the holes for inlet and outlet connections, the PDMS chip was thermally aged at 200  $^{\circ}\text{C}$  for 12 h to reduce pressure-induced deformation<sup>58</sup>, yielding a chip with a Young's modulus of<sup>59</sup> approximately 3 MPa. Both the PDMS chip and the glass substrate were cleaned with isopropanol and treated by plasma for 90 s before bringing them into contact. Once the PDMS chip was sealed against the glass slide, the device was placed in an oven for 30 min at 74  $^{\circ}\text{C}$  to improve bonding quality.

The mould used was a silicon wafer containing microchannel patterns created by soft photolithography using a negative photoresist<sup>60,61</sup>. A 4-inch silicon wafer (test grade, University Wafer, Boston, MA) was cleaned with acetone and isopropanol and dried with nitrogen gas. The wafer was then coated with SU-8 50 negative photoresist (MicroChem Corp., Newton, MA) on a spin coater (Laurell Technologies Corp., North Wales, PA) operating at 600 rpm for 30 s. After a pre-exposure bake at 65  $^{\circ}\text{C}$  and subsequently at 95  $^{\circ}\text{C}$ , each for 60 min, the coated wafer was exposed to UV light (Autoflood 1000, Optical Associates, Milpitas, CA) through a negative transparent photomask that contained the desired channel design. Following a 3.5 min post-exposure bake at 95  $^{\circ}\text{C}$ , the wafer was developed in SU-8 developer (MicroChem Corp., Newton, MA) for 60 min to obtain the pattern.

## Flexdym channel fabrication

Flexdym (Blackholelab Inc., Paris) is a thermoplastic elastomer (Young's modulus of 1.18 MPa) with a rapid and easy moulding process for microfluidic devices<sup>62</sup>. After fabrication of the silicon wafer mould containing the channel designs, a sheet of Flexdym (6 cm  $\times$  4 cm) was placed directly above the mould with another sheet of unpatterned PDMS (about 1 mm thick) placed above the Flexdym for protection. The whole set was then placed on a heat press between two Teflon sheets. The plate on the heat press was heated to 175  $^{\circ}\text{C}$  before starting to mould the Flexdym. Once the target temperature was reached, the lever on the heat plate was locked down with a timer set for 5 min. After the process was finished, the lever was released and the Flexdym sheet was inspected visually to make sure that no bubbles were trapped around the channel. The chip was allowed to cool down for 5 min before unfolding the layers. The Flexdym was permanently sealed with a glass slide by following the same sealing procedure used for the PDMS channels. The dimensions of the cross-section of the channels were 201  $\mu\text{m}$  (width)  $\times$  166  $\mu\text{m}$  (height), and the diameter of the obstacles was 99  $\mu\text{m}$ .

## SU-8 photoresist channel fabrication

To make microfluidic channels directly from SU-8 photoresist, an inverse mask was designed and printed on transparency. The desired channel was printed on the inverse mask in black with transparent dots marking the obstacles, and the rest of the mask was left transparent. The same procedure used to make the silicon wafer master as described in Methods section 'PDMS channel fabrication' was followed to fabricate the channels on glass slides. The chip was then sealed by 3M VHB tape to another glass slide with holes for connections. The dimensions of the cross-section of the channels were 209  $\mu\text{m}$  (width)  $\times$  196  $\mu\text{m}$  (height), and the diameter of the obstacles was 90  $\mu\text{m}$ . The Young's modulus of SU-8 photoresist is 2 GPa (from the table of properties for SU-8 permanent photoresists, MicroChem Corp., Newton, MA, available at <http://microchem.com/pdf/SU-8-table-of-properties.pdf>).

## Flow rate measurement

Experimental measurements in Figs. 2 and 4 were made with the system shown in Fig. 4a. When measuring the relation between pressure and flow rate, the linking channel valve was closed to allow separate measurement of the channel with and the channel without obstacles. Deionized (DI) water was pumped through each channel and a pressure scan from 0 to 100 kPa was performed using an Elveflow OB1 pressure controller. The flow rate was measured by an Elveflow MFS5 flow sensor (0.2–5 ml min<sup>-1</sup>). To verify Braess's paradox, the same instruments were used and the pressure was set constant while recording the flow

rate at each outlet. Red (3 g l<sup>-1</sup>, FD&C Red #40, Flavors & Colours) and blue (1.5 g l<sup>-1</sup>, FD&C Blue #1, Flavors & Colours) dyes were added into DI water to demonstrate the switching behaviour. The concentrations of the dyes were adjusted for similar flow rate under the same pressure. The flow rate measurements in Supplementary Fig. 10 were performed using isolated channels constructed from Flexdym and SU-8 photore-sist, respectively.

## Fluorescence imaging

Fluorescent polyethylene microspheres (10–20  $\mu\text{m}$ ) were suspended in Tween 80 solution (Cospheric LLC, Santa Barbara, CA) and pumped through a single microfluidic channel with obstacles by an Elveflow OB1 pressure controller. Two different pressures were applied, 3 kPa and 100 kPa, to demonstrate different flow profiles around the obstacles. Fluorescence images were captured with an Olympus BX51 microscope equipped with a NIBA filter through an Infinity 3 CCD camera.

## Measured flow rate data and statistics

Savitsky–Golay filtering was applied to all flow rate data collected through experiments, using a window length of 11 data points and a second-order polynomial. For each of the fixed pressures presented in Fig. 4b–d, a 60 s time series of flow rate data was collected at each of the outlets with a sampling rate of 10 Hz. Over the 60 s interval, the linking channel valve was sequentially opened/closed every 15 s. For each time series, the 15 s intervals in which the valve was open (closed) were averaged to create a single 15 s time series for each outlet. The total flow rate ( $Q_1 + Q_2$ ) was calculated when the valve is open and closed, respectively, by summing the 15 s time series for the two outlets point-by-point. The statistics presented in Fig. 4 are the average and standard deviation of the resulting series. For Supplementary Fig. 12, the flow rate at each of the two outlets was measured experimentally at a sampling rate of 100 Hz over a 180 s interval, during which the linking channel was sequentially opened/closed every 30 s. The total flow rate in Supplementary Fig. 12c was calculated by summing, point-by-point, the data in Supplementary Fig. 12a, b.

## Parameters in simulations and experiments

In the simulations, we set  $\rho = 10^3 \text{ kg m}^{-3}$ ,  $\mu = 10^{-3} \text{ Pa s}$ ,  $\nu = \mu/\rho = 10^{-6} \text{ m}^2 \text{ s}^{-1}$ ,  $w = 500 \text{ }\mu\text{m}$  for the width of all channels, and  $r = 100 \text{ }\mu\text{m}$  for the radius of all obstacles, unless otherwise noted. In all experiments, DI water was used as the working fluid. The other undeclared dimensions were as follows. In Fig. 2a, b, the length of the (partially shown) channel was 1.25 cm. In Fig. 2c–e, the channel length was 4.3 cm, and in Fig. 2f the channel length was 2.0 cm (see Methods section 'PDMS channel fabrication' for the remaining dimensions). In Fig. 3,  $L_1 = 0.17 \text{ cm}$ ,  $L_2 = 1.0 \text{ cm}$ ,  $L_3 = 0.1 \text{ cm}$ ,  $L_4 = 1.25 \text{ cm}$ , and  $L_5 = 1.0 \text{ cm}$ . In Fig. 4,  $L_1 = 0.6 \text{ cm}$ ,  $L_2 = 2.9 \text{ cm}$ ,  $L_4 = 1.4 \text{ cm}$  and  $L_5 = 1.4 \text{ cm}$ . For the linking channel, the switch valve was connected to the two parallel channels through 15 cm of round tubing and 0.7 cm of microchannel on each side. Each inlet was connected to the pressurized vials through 62 cm of tubing, and each outlet was attached to 50 cm of tubing. The inner diameter of all tubing was 0.79 mm.

## Data availability

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

## Code availability

Custom Python code is available from the corresponding author on request.

48. OpenFOAM v4.1 (OpenFOAM Foundation, 2016).

49. Geuzaine, C. & Remacle, J.-F. Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities. *Int. J. Numer. Methods Eng.* **79**, 1309–1331 (2009).

50. Oh, K. W., Lee, K., Ahn, B. & Furlani, E. P. Design of pressure-driven microfluidic networks using electric circuit analogy. *Lab Chip* **12**, 515–545 (2012).
51. Zeitoun, R. I., Langelier, S. M. & Gill, R. T. Implications of variable fluid resistance caused by start-up flow in microfluidic networks. *Microfluid. Nanofluidics* **16**, 473–482 (2014).
52. Zovatto, L. & Pedrizzetti, G. Flow about a circular cylinder between parallel walls. *J. Fluid Mech.* **440**, 1–25 (2001).
53. Gervais, T., El-ali, J., Gunther, A. & Jensen, K. F. Flow-induced deformation of shallow microfluidic channels. *Lab Chip* **6**, 500–507 (2006).
54. Christov, I. C., Cognet, V., Shidhore, T. C. & Stone, H. A. Flow rate–pressure drop relation for deformable shallow microfluidic channels. *J. Fluid Mech.* **841**, 267–286 (2018).
55. Amstad, E., Datta, S. S. & Weitz, D. A. The microfluidic post-array device: high throughput production of single emulsion drops. *Lab Chip* **14**, 705–709 (2014).
56. Haudin, F., Callewaert, M., De Malsche, W. & De Wit, A. Influence of nonideal mixing properties on viscous fingering in micropillar array columns. *Phys. Rev. Fluids* **1**, 074001 (2016).
57. Zhao, H., Liu, Z., Zhang, C., Guan, N. & Zhao, H. Pressure drop and friction factor of a rectangular channel with staggered mini pin fins of different shapes. *Exp. Therm. Fluid Sci.* **71**, 57–69 (2016).
58. Kim, M., Huang, Y., Choi, K. & Hidrovo, C. H. The improved resistance of PDMS to pressure-induced deformation and chemical solvent swelling for microfluidic devices. *Microelectron. Eng.* **124**, 66–75 (2014).
59. Johnston, I. D., McCluskey, D. K., Tan, C. K. L. & Tracey, M. C. Mechanical characterization of bulk sylgard 184 for microfluidics and microengineering. *J. Micromech. Microeng.* **24**, 035017 (2014).
60. Martin, R. S., Gawron, A. J., Lunte, S. M. & Henry, C. S. Dual-electrode electrochemical detection for poly(dimethylsiloxane)-fabricated capillary electrophoresis microchips. *Anal. Chem.* **72**, 3196–3202 (2000).
61. Duffy, D. C., McDonald, J. C., Schueller, O. J. A. & Whitesides, G. M. Rapid prototyping of microfluidic systems in poly(dimethylsiloxane). *Anal. Chem.* **70**, 4974–4984 (1998).
62. Lachaux, J. et al. Thermoplastic elastomer with advanced hydrophilization and bonding performances for rapid (30 s) and easy molding of microfluidic devices. *Lab Chip* **17**, 2581–2594 (2017).

**Acknowledgements** This research was supported by the US National Science Foundation (grants PHY-1001198 and CHE-1900011), the Simons Foundation (award number 342906) and a Northwestern University Presidential Fellowship.

**Author contributions** D.J.C., J.-R.A. and A.E.M. designed the overall study and formulated the theory. Y.L. and I.Z.K. designed and performed the experiments. D.J.C. implemented the numerical simulations and analyses. All authors contributed to the writing of the manuscript, which was led by D.J.C. and A.E.M. All authors reviewed and approved the final manuscript.

**Competing interests** The authors declare no competing interests.

**Additional information**

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1701-6>.

**Correspondence and requests for materials** should be addressed to A.E.M.

**Peer review information** *Nature* thanks Sujit Datta and Dino Di Carlo for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

# Superconductors, orbital magnets and correlated states in magic-angle bilayer graphene

<https://doi.org/10.1038/s41586-019-1695-0>

Received: 15 March 2019

Accepted: 12 August 2019

Published online: 30 October 2019

Xiaobo Lu<sup>1</sup>, Petr Stepanov<sup>1</sup>, Wei Yang<sup>1</sup>, Ming Xie<sup>2</sup>, Mohammed Ali Aamir<sup>1</sup>, Ipsita Das<sup>1</sup>, Carles Urgell<sup>1</sup>, Kenji Watanabe<sup>3</sup>, Takashi Taniguchi<sup>3</sup>, Guangyu Zhang<sup>4</sup>, Adrian Bachtold<sup>1</sup>, Allan H. MacDonald<sup>2</sup> & Dmitri K. Efetov<sup>1\*</sup>

Superconductivity can occur under conditions approaching broken-symmetry parent states<sup>1</sup>. In bilayer graphene, the twisting of one layer with respect to the other at ‘magic’ twist angles of around 1 degree leads to the emergence of ultra-flat moiré superlattice minibands. Such bands are a rich and highly tunable source of strong-correlation physics<sup>2–5</sup>, notably superconductivity, which emerges close to interaction-induced insulating states<sup>6,7</sup>. Here we report the fabrication of magic-angle twisted bilayer graphene devices with highly uniform twist angles. The reduction in twist-angle disorder reveals the presence of insulating states at all integer occupancies of the fourfold spin–valley degenerate flat conduction and valence bands—that is, at moiré band filling factors  $\nu = 0, \pm 1, \pm 2, \pm 3$ . At  $\nu \approx -2$ , superconductivity is observed below critical temperatures of up to 3 kelvin. We also observe three new superconducting domes at much lower temperatures, close to the  $\nu = 0$  and  $\nu = \pm 1$  insulating states. Notably, at  $\nu = \pm 1$  we find states with non-zero Chern numbers. For  $\nu = -1$  the insulating state exhibits a sharp hysteretic resistance enhancement when a perpendicular magnetic field greater than 3.6 tesla is applied, which is consistent with a field-driven phase transition. Our study shows that broken-symmetry states, interaction-driven insulators, orbital magnets, states with non-zero Chern numbers and superconducting domes occur frequently across a wide range of moiré flat band fillings, including close to charge neutrality. This study provides a more detailed view of the phenomenology of magic-angle twisted bilayer graphene, adding to our evolving understanding of its emergent properties.

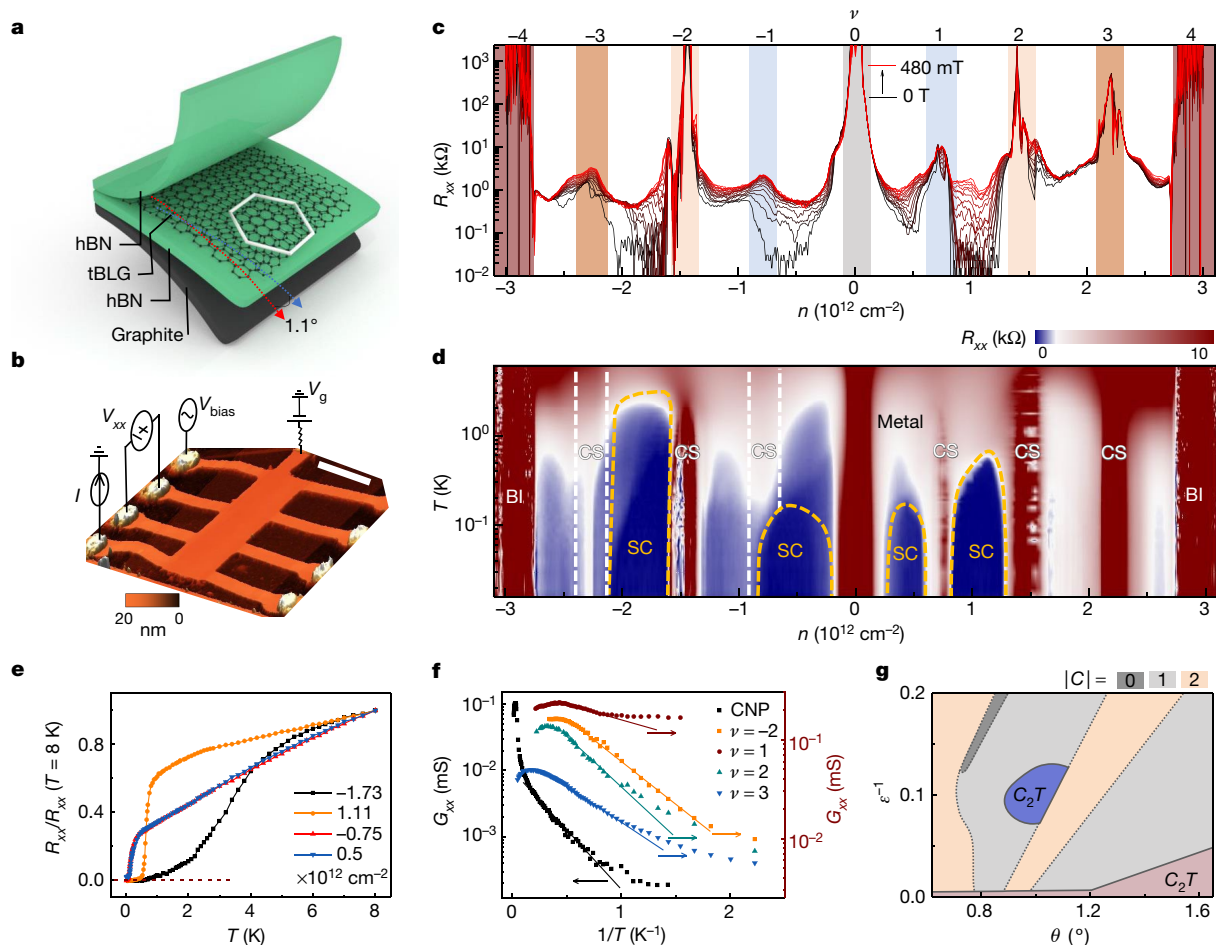
Interactions dominate over single-particle physics in flat-band electronic systems, and can give rise to insulating states at partial band fillings<sup>3,4</sup>, superconductivity<sup>8</sup> and magnetism<sup>9–14</sup>. Recently, correlated insulating phases and strongly coupled superconducting domes have been found in ultra-flat bands of magic-angle twisted bilayer graphene (MAG) close to half-filling ( $\nu = \pm 2$ ), establishing graphene as a platform for the investigation of strongly correlated two-dimensional electrons<sup>6,15–18</sup>. MAG has several advantages that should enable new insights into these systems: the correlations can be accurately controlled by varying the twist angle between the two graphene layers; techniques for the fabrication of ultra-clean graphene layers are well-established; and the electron density ( $n_0 = A_0^{-1} \approx 10^{12} \text{ cm}^{-2}$ , where  $A_0$  is the area of the moiré unit cell) that is required to fill a moiré superlattice band can be adequately supplied by electrical gates.

Here we report the observation of correlated states at all integer fillings of  $\nu = n/n_0$  (where  $n$  is the gate-modulated carrier density),

including at charge neutrality, and the occurrence of new superconducting domes and orbital magnetic states in MAG. When interactions are neglected, the two low-energy moiré bands of MAG have fourfold spin–valley flavour degeneracies, which implies that the density measured from the carrier neutrality point (CNP) is  $4n_0$  when the flat conduction band is full and  $-4n_0$  when the valence band is empty<sup>2,19</sup>. Interactions can lift the flavour degeneracies and give rise to completely empty or full spin–valley polarized flat bands—with interaction-induced gaps at all integer values of  $\nu$ —in place of the symmetry-protected Dirac points that connect the conduction and valence bands for each flavour<sup>10</sup>. The many-body physics of these bands is highly sensitive to the twist angle  $\theta$  and the interaction strength  $\varepsilon^{-1}$  (where  $\varepsilon$  is the effective dielectric constant in MAG). In some cases—depending on the details of the electronic structure—bands can have non-zero Chern numbers<sup>9,10,20–23</sup>, allowing for the possibility of orbital magnetism and anomalous Hall effects.

<sup>1</sup>ICFO – Institut de Ciències Fotoniques, The Barcelona Institute of Science and Technology, Castelldefels, Barcelona, Spain. <sup>2</sup>Department of Physics, University of Texas at Austin, Austin, TX, USA. <sup>3</sup>National Institute for Materials Science, Tsukuba, Japan. <sup>4</sup>Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing, China. \*e-mail: dmitri.efetov@icfo.eu





**Fig. 1 | Integer-filling correlated states and new superconducting domes.**

**a**, Schematic of a typical MAG device. **b**, Atomic force microscopy image and schematic of how various measurements are obtained. Scale bar, 2  $\mu\text{m}$ . **c**, Four-terminal longitudinal resistance plotted against carrier density at different perpendicular magnetic fields from 0 T (black trace) to 480 mT (red trace). **d**, Colour plot of longitudinal resistance against carrier density and temperature, showing different phases including metal, band insulator (BI), correlated state (CS) and superconducting state (SC). The boundaries of the superconducting domes—indicated by yellow lines—are defined by 50% resistance values relative to the normal state. Note that the transition from the metal to the superconducting state is not sharp at some carrier densities, which adds uncertainty to the value of  $T_c$  extracted. **e**, Longitudinal resistance at optimal doping of the superconducting domes as a function of temperature. The resistance is normalized to its value at 8 K. Note that data points for

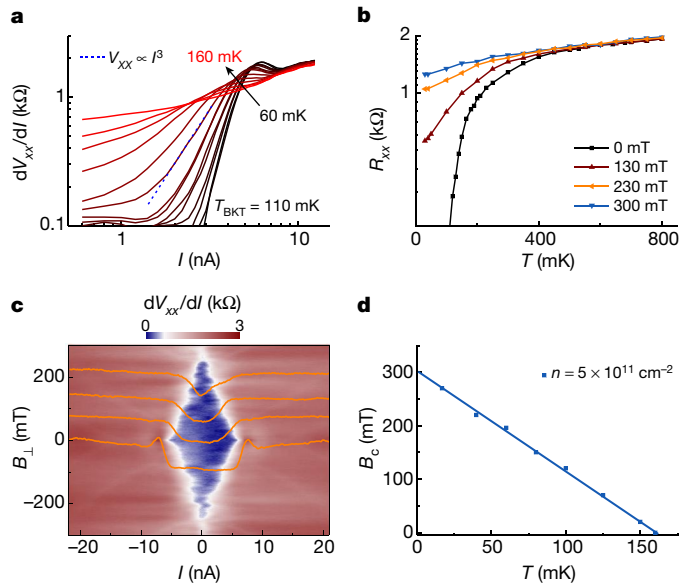
$n = -7.5 \times 10^{11} \text{ cm}^{-2}$  are overlaid by the data points for  $n = 5 \times 10^{11} \text{ cm}^{-2}$ , as both curves follow a very similar line. **f**, Conductance  $G_{xx}$  plotted against inverse temperature at carrier densities corresponding to  $\nu = 0, 1, \pm 2$  and 3. The straight lines are fits to  $G_{xx} \propto \exp(-\Delta/2kT)$  (where  $\Delta$  is the size of the correlation-induced gap and  $k$  is the Boltzmann constant), for temperature-activated behaviour, and give gap values of 0.35 meV ( $\nu = -2$ ), 0.14 meV ( $\nu = 1$ ), 0.37 meV ( $\nu = 2$ ), 0.27 meV ( $\nu = 3$ ) and 0.86 meV ( $\nu = 0$ ; CNP). **g**, Mean-field phase diagram for neutral  $\nu = 0$  (CNP) twisted bilayer graphene, as a function of twist angle and interaction strength, showing different configurations of  $C_2T$  symmetry and Chern number ( $C$ ). Red and blue regions with solid outlines indicate states that do not break symmetry, and therefore have bands with no Berry curvature and vanishing Chern number. Blue indicates a gapped state and red indicates a gapless state. Zones filled with other colours indicate gapped states that break  $C_2T$  symmetry and have bands with different Chern numbers, as shown.

Figure 1a is a schematic of a typical graphite-gated, hexagonal boron nitride (hBN)-encapsulated MAG heterostructure device. The atomic force microscopy image in Fig. 1b shows the high structural homogeneity of the device. Figure 1c shows four-terminal resistance  $R_{xx}$  as a function of  $n$  at different out-of-plane magnetic fields  $B_{\perp}$ , measured at a temperature  $T$  of 16 mK. We find strong resistance peaks at  $n = 4n_0 \approx \pm 3 \times 10^{12} \text{ cm}^{-2}$  that mark the edges of the flat bands, consistent with previous studies<sup>3,6,18</sup>. The full-band density corresponds to an average twist angle across the device of about  $1.10^\circ$ . By comparing  $2n_0$  values extracted from two-terminal measurements between different contact pairs (Extended Data Fig. 4), we estimate that the variation in twist angle ( $\Delta\theta$ ) is only around  $0.02^\circ$  over a span of about  $10 \mu\text{m}$ . Such homogeneity in the twist angle is, to our knowledge, unprecedented in a MAG device.

In addition to the resistance peaks at the CNP and at  $\nu = \pm 4$ , we also observe interaction-induced resistance peaks at all non-zero integer fillings of the moiré bands ( $\nu = \pm 1, \pm 2, \pm 3$ ), corresponding to 1, 2 and

3 electrons (+) or holes (−) per moiré unit cell (Fig. 1c). Signatures of some of these resistive states have been observed previously<sup>3,6,18,24</sup>, but they are much more strongly developed here. From temperature-dependent transport behaviour over a range of 10 K (Fig. 1f), it is possible to extract the activated gap size of the correlated insulator states. We obtain values of 0.34 meV ( $\nu = -2$ ), 0.37 meV ( $\nu = 2$ ) and 0.25 meV ( $\nu = 3$ ). Evidence for thermally activated transport is much weaker for the  $\nu = 1$  state (0.14 meV) and is entirely absent for the  $\nu = -3$  and  $\nu = -1$  states, which might indicate that these are correlated semi-metallic states rather than insulating states<sup>25</sup>.

Our device also shows clear temperature-activated transport behaviour below 33 K at the CNP, with an extracted gap size of 0.86 meV. Gaps at the CNP do not require broken flavour symmetries, but they do require that at least one of the emergent  $C_3$  and  $C_2T$  symmetries—which prevent CNP bands from touching—be broken. These symmetries can be explicitly broken by crystallographic alignment of the MAG and hBN



**Fig. 2 | The superconducting dome at fillings between  $\nu = 0$  and  $\nu = 1$ .**

**a**, Differential resistance plotted against d.c. bias current at various temperatures from 60 mK (black trace) to 160 mK (red trace). The blue dashed line is a fit to the  $V_{xx} \approx I^3$  power law, and identifies a BKT transition at a temperature ( $T_{\text{BKT}}$ ) of around 110 mK. **b**, Longitudinal resistance plotted against temperature at various out-of-plane magnetic fields, showing that normal levels of resistance are restored at magnetic fields greater than 300 mT. **c**, Two-dimensional colour plot of the differential resistance as a function of magnetic field and excitation current at 16 mK. The orange traces show differential resistance plotted against current at magnetic field values of 225 mT, 150 mT, 75 mT and 0 T (top to bottom). **d**, Values of the critical magnetic field at various temperatures. The straight line is a fit to the Ginzburg–Landau expression. For all measurements in **a–d**, the carrier density was fixed at optimal doping of the dome  $n = 5 \times 10^{11} \text{ cm}^{-2}$ .

layers; however, careful inspection of the angle between these (see Supplementary Information) allows us to rule out this scenario. As we also do not observe any other signatures of hBN alignment, such as satellite resistance peaks<sup>9</sup>, we conclude that the gap at CNP probably originates as a result of interactions.

The existence of a non-trivial gap at the CNP has strong implications for the properties of other gapped MAG states. Mean-field theory (Fig. 1g, Supplementary Information), predicts gapped states at neutrality over a wide range of twist angles and interaction strengths. Gapped states at non-zero integer values of  $\nu$  are expected only when the moiré superlattice band width is smaller than the exchange shift produced by band occupation, and this occurs only near the magic angle. Overall our calculations demonstrate that insulating—or for weak interactions, semi-metallic states—are common at all integer values of  $\nu$ , as observed experimentally. This mean-field phase diagram does not allow for broken translational symmetry, which appears not to be required for our experiments. If broken translational symmetry did have a key role in establishing insulating states, they would be expected at moiré band fillings  $\nu = n + p/3$  where  $p$  and  $n$  are integers; this is not consistent with our experimental observations.

Notably, in four distinct carrier-density intervals between integer filling factors, we observe sharp decreases in the resistance of the device with decreasing temperature (Fig. 1e), which can be restored by the application of a small perpendicular magnetic field  $B_{\perp} < 500 \text{ mT}$  (Fig. 2b). Figure 1d shows a colour plot of resistance against temperature and carrier density, in which four dome-shaped pockets of low resistance flank the most resistive states. In three of these domes, the resistance decreases to zero (Fig. 1e, Extended Data Figs. 5, 7), which is consistent with superconductivity. In the fourth region (at  $n = 5 \times 10^{11} \text{ cm}^{-2}$ )

the resistance remains slightly greater than zero, owing to insufficient cooling of electrons below 100 mK in the cryostat<sup>7,18</sup>.

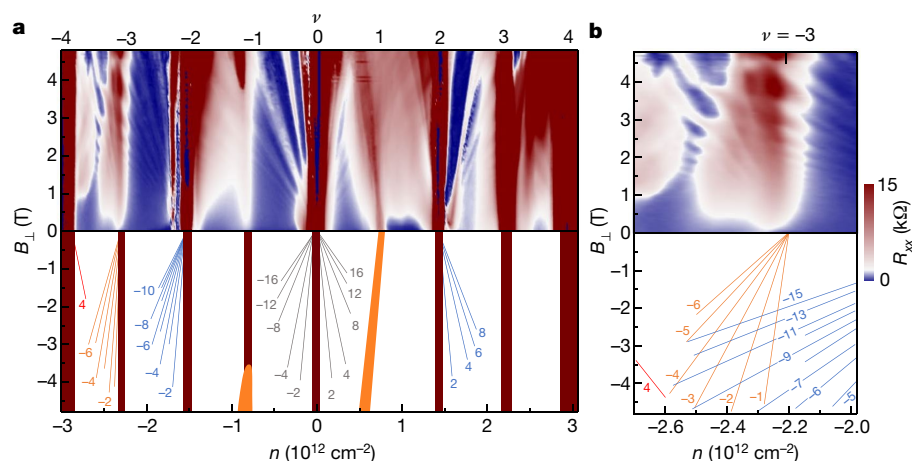
In the dome close to  $-2n_0$ , we observe a superconducting transition. Although this has been reported previously, the superconducting transition temperature  $T_c$  (defined as half the normal-state resistance) of 3 K that we observe here is considerably higher than the previous value<sup>6</sup> of around 1.7 K. The other superconducting domes, which to our knowledge are all observed here for the first time, have much lower  $T_c$  values and much sharper transitions. We identify a superconducting dome between  $n_0$  and  $2n_0$  with  $T_c \approx 650 \text{ mK}$ , and two domes between the CNP and  $\pm n_0$  with  $T_c \approx 160 \text{ mK}$  and  $T_c \approx 140 \text{ mK}$ , respectively. As can be seen in Fig. 1d and Extended Data Fig. 6, it is likely that additional superconducting domes are developing between other filling factors; however, these are not fully developed, and are presumably obscured by the inhomogeneity that remains in our improved samples.

Figure 2 shows the signatures of the newly observed superconducting domes (exemplified by the state between CNP and  $n_0$ ; all other states are described in detail in Extended Data Fig. 7 and Extended Data Table 1). In Fig. 2a, the differential resistance  $dV_{xx}/dI$  is plotted against the d.c. bias current  $I$  at various temperatures. At 60 mK, the traces display the nonlinear resistance typical of two-dimensional superconductivity, with a sharp resistive transition for  $I > I_c \approx 3 \text{ nA}$  (where  $I_c$  is the critical supercurrent). The blue dashed line is a power-law fit to  $dV_{xx}/dI \approx I^2$ , consistent with two-dimensional superconductivity described by the Berezinskii–Kosterlitz–Thouless (BKT) theory, showing a transition temperature  $T_{\text{BKT}}$  of around 110 mK.

The temperature dependence of the resistance  $R_{xx}$  at various magnetic fields is illustrated in Fig. 2b. The superconductivity signal is gradually weakened upon increasing the applied field, and  $R_{xx}$  varies almost linearly with temperature above a critical field  $B_c \approx 300 \text{ mT}$ . The suppression of superconductivity by the magnetic field is further exemplified by Fig. 2c, which shows a plot of the differential resistance as a function of the magnetic field and the excitation current  $I$  at 16 mK. The critical supercurrent  $I_c$  is reduced by application of a magnetic field, reaching zero when  $B_c > 300 \text{ mT}$ . From these measurements, we extract the temperature-dependent critical magnetic field  $B_c$  (defined by 50% of the normal state  $R_{xx}$  value). By fitting to the expression from Ginzburg–Landau theory,  $B_c = [\Phi_0/(2\pi\xi^2)](1 - T/T_c)$ , we extract a coherence length  $\xi_{\text{GL}}(T=0 \text{ K})$  of around 32 nm. Here  $\Phi_0 = h/(2e)$  is the superconducting flux quantum,  $h$  is Planck’s constant and  $e$  is the electron charge.

We have studied the response of the flat bands to an applied magnetic field at a temperature of 100 mK. Figure 3a shows a colour map of the resistance as a function of carrier density and magnetic field, and the corresponding schematic highlights the trajectories of the resistance maxima. We find sets of Landau fans that originate from the CNP and from most of the resistive states with an integer filling factor. In previous studies, Landau levels were identified only on the high-carrier-density sides of insulating states<sup>3,18</sup>. Here, we also observe Landau levels dispersing to lower densities. The vanishing carrier densities near most integer filling factors—as evidenced by both Landau fans and weak field Hall resistivities (Extended Data Fig. 3)—suggest that the fourfold spin–valley band degeneracy of the non-interacting state is lifted over a large range of filling factors, resetting the carrier density per band.

Our observations suggest that a rich variety of spin–valley broken-symmetry states occur as a function of carrier density and magnetic field. Landau levels that can be traced to the CNP exhibit fourfold degeneracy with a filling-factor sequence of  $\nu_l = \pm 4, \pm 8, \pm 12, \dots$ , as well as spin–valley broken-symmetry states with  $\nu_l = \pm 2$ . The Landau levels that fan out from  $\nu = 2(-2)$  follow a sequence of  $\nu_l = 2(-2), 4(-4), 6(-6), \dots$  at low magnetic field, indicating partially lifted degeneracy for either spin or valley. Quantum oscillations from  $\nu = -2$  exhibit a dominant degeneracy sequence of  $\nu_l = -3, -5, -7, \dots$  at high magnetic field. Near  $\nu = -3$ , quantum oscillations exhibit fully lifted degeneracy of Landau levels with filling factors  $\nu_l = -1, -2, -3, -4, \dots$ . The Landau fans that emerge from insulating



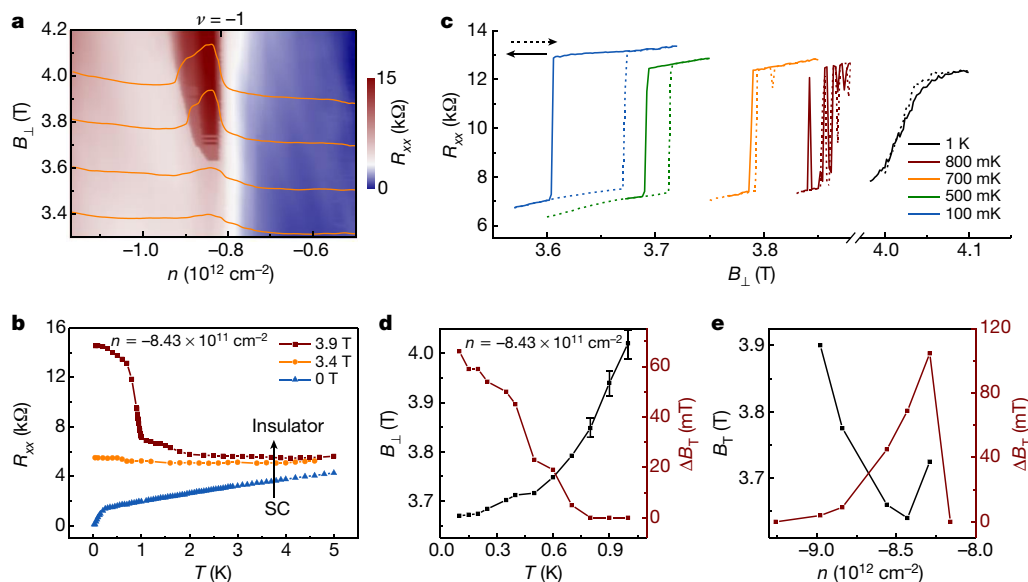
**Fig. 3 | Shubnikov-de Haas oscillations in the MAG flat bands.** **a**, Top, colour map of longitudinal resistance plotted against carrier density and magnetic field. Bottom, the corresponding schematic that identifies visible Landau level fans with a dominant degeneracy. The Landau fan diagram diverging from the CNP ( $\nu=0$ ) follows a fourfold degenerate sequence with  $\nu_L = \pm 4, \pm 8, \pm 12, \dots$ , with symmetry-broken states at  $\nu_L = \pm 2$ . The fans from  $\nu=2(-2)$  follow a twofold degenerate  $\nu_L = -2(2), -4(4), -6(6), \dots$  sequence, with broken-symmetry states at

$\nu_L = -3, -5, -7, \dots$ . The  $\nu=-3$  fan follows a single degenerate  $\nu_L = -1, -2, -3, -4, \dots$  sequence. Emergent correlated phases at all integer moiré fillings, including the CNP ( $\nu=0$ ), are highlighted in dark red. Chern insulating states are highlighted in orange. **b**, Magnification of **a** around the  $\nu=-3$  state, showing signatures similar to a Hofstadter butterfly spectrum with criss-crossing Landau levels fanning out from  $\nu=-3$  and  $\nu=-2$  filling states.

states all extrapolate to a carrier density that vanishes at integer moiré band filling factors.

We also find that the degeneracies of Landau levels originating from the CNP and  $\nu=-2$  change when crossing the  $\nu=-1$  and  $\nu=-3$  states, suggestive of first-order phase transitions that change band degeneracies. In particular, as is shown in Fig. 3b, Landau levels from the  $\nu=-2$  and  $\nu=-3$  states display a criss-crossing pattern—superficially similar to that of a Hofstadter butterfly, but distinct in that the Landau level indices that can be traced to the  $\nu=-3$  state are spaced by one filling, whereas those that can be traced to the  $\nu=-2$  state are spaced by two fillings.

It is noteworthy that neither of the  $\nu=\pm 1$  correlated states show clear formation of Landau levels. The positions of their resistance maxima do, however, exhibit clear dependencies on the magnetic field and the carrier density. At  $\nu=-1$  the resistance state has no slope ( $dn/dB$ ) at low field; however, above a critical field  $B_T \approx 3.6$  T, we observe the sudden development of a slope that is consistent with a Chern number of 1. Furthermore, at  $\nu=1$  the position of the resistance peak shifts to lower carrier density, with a slope that is consistent with a Chern number of 2. The slope of  $dn/dB$  in the absence of a Landau-level fan in the  $\nu=\pm 1$  correlated states is consistent with Chern insulating states from spin and valley symmetry breaking at odd values of  $\nu$ . As discussed earlier



**Fig. 4 | Field-driven phase transition near the  $\nu=-1$  state.** **a**, Longitudinal resistance plotted as a function of carrier density and out-of-plane magnetic field measured at 16 mK. The orange traces show longitudinal resistance plotted against carrier density with the magnetic field (from top to bottom) fixed at 4 T, 3.8 T, 3.6 T and 3.4 T. **b**, Longitudinal resistance plotted against temperature at various magnetic fields. **c**, Longitudinal resistance plotted against magnetic field at various temperatures, with dashed and solid lines corresponding to

increasing and decreasing magnetic field, respectively. **d**, Dependence of the critical magnetic field (extracted from up sweeps) and the hysteresis value on temperature. Note that the transition above 800 mK is not sharp, adding uncertainty to the extracted critical field values. In **b–d**, the carrier density is fixed at  $-8.43 \times 10^{11} \text{ cm}^{-2}$ . **e**, Dependence of the critical magnetic field and the hysteresis value on the carrier density at 100 mK.



and as predicted by mean-field theory, valley-projected bands in insulating states can have non-zero Chern numbers that compete closely with states with zero Chern numbers. Although we cannot resolve quantized values in  $R_{xy}$  nor zero resistance in  $R_{xx}$ , as expected for a Chern insulating state, we do not do so for the other Landau levels in Fig. 3a either. We therefore conclude that our devices are still too inhomogeneous to observe quantization over the entire device.

Exactly at the transition at which the slope of the  $\nu = -1$  resistive state in Fig. 3a changes from  $dn/dB = 0$  to a  $dn/dB$  consistent with Chern number 1, we find a strong hysteretic increase of  $R_{xx}$  and  $R_{xy}$ ; this is indicative of a possible magnetic-field-induced first-order phase transition. Figure 4a displays a plot of  $R_{xx}$  as a function of  $n$  and  $B_{\perp}$ . Figure 4b displays the temperature-dependent resistance  $R_{xx}(T)$  near  $\nu = -1$  (or  $n = -8.43 \times 10^{11} \text{ cm}^{-2}$ ) for a series of magnetic field values. Whereas at  $B_{\perp} = 0 \text{ T}$ ,  $R_{xx}(T)$  shows a typical metal-superconductor phase transition, above  $B_{\perp} > 3.6 \text{ T}$  and below  $T < 0.9 \text{ K}$ ,  $R_{xx}(T)$  has a sharp jump and an insulating temperature dependence.

Figure 4c shows plots of  $R_{xx}$  against the magnetic field at  $\nu = -1$  for up and down sweeps of the magnetic field. Below  $800 \text{ mK}$ , the curves show sharp jumps in resistance at associated critical transition fields  $B_T$ , and demonstrate strong hysteretic behaviour that is dependent on the sweeping direction of the magnetic field; the width of the magnetic field of the hysteresis loop is denoted by  $\Delta B_T$ . The critical field  $B_T$  is always higher for up sweeps than for down sweeps.

Both  $B_T$  and  $\Delta B_T$  are highly temperature-dependent, with  $B_T$  shifting to higher values and  $\Delta B_T$  becoming smaller as the temperature increases. At  $T > 800 \text{ mK}$ , the hysteresis almost disappears and the transition becomes broader. The temperature dependencies of  $B_T$  and  $\Delta B_T$  were extracted and are shown in Fig. 4d. The phase transition and hysteresis occur over a narrow range of carrier densities from around  $-8.3 \times 10^{11} \text{ cm}^{-2}$  to  $-9 \times 10^{11} \text{ cm}^{-2}$  (Extended Data Fig. 8b, c) with  $B_T$  and  $\Delta B_T$  at different carrier densities shown in Fig. 4e. Overall, we observe similar behaviour in Hall resistance measurements (Extended Data Fig. 8d). These observations indicate that the origin of the change in the slope  $dn/dB$  of the resistance maximum is a first-order phase transition, and is probably due to a competition between correlated states with zero and non-zero Chern numbers at high magnetic fields<sup>26</sup>, suggesting the emergence of a field-stabilized orbital magnetic state.

Notably, we have observed superconducting domes close to charge neutrality. To our knowledge, these states represent the lowest carrier density ( $n \approx 3 \times 10^{11} \text{ cm}^{-2}$ ; counting from CNP) at which superconductivity has been observed. The existence of superconducting domes across a wide range of moiré band fillings must have important implications for our understanding of their origin. Because the density of states diminishes close to the CNP, the appearance of superconductivity seems not to be simply related to a high density of states of the non-interacting bands. Superconductivity occurs adjacent to insulating states that seem—on the basis of Landau fan patterns—to break spin–valley degeneracy, and adjacent to insulating states that do not. Nevertheless, its consistent association with nearby correlated insulator states suggests an exotic pairing mechanism. Conversely, at this point our observation cannot rule out the possibility of conventional electron–phonon coupling superconductivity in metallic states with quasiparticles that evolve adiabatically from those of the non-interacting system and compete with a rich variety of distinct insulating states from which they are separated by first-order phase transition lines<sup>24,27,28</sup>. In this case, it is possible that the consistent high density of states over a broad range of filling factors helps to support superconductivity in the metallic state.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1695-0>.

- Lee, P. A., Nagaosa, N. & Wen, X.-G. Doping a Mott insulator: physics of high-temperature superconductivity. *Rev. Mod. Phys.* **78**, 17 (2006).
- Bistritzer, R. & MacDonald, A. H. Moiré bands in twisted double-layer graphene. *Proc. Natl Acad. Sci. USA* **108**, 12233–12237 (2011).
- Cao, Y. et al. Correlated insulator behaviour at half-filling in magic-angle graphene superlattices. *Nature* **556**, 80–84 (2018).
- Chen, G. et al. Evidence of a gate-tunable Mott insulator in a trilayer graphene moiré superlattice. *Nat. Phys.* **15**, 237–241 (2019).
- Tarnopolsky, G., Kruchkov, A. J. & Vishwanath, A. Origin of magic angles in twisted bilayer graphene. *Phys. Rev. Lett.* **122**, 106405 (2019).
- Cao, Y. et al. Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).
- Chen, G. et al. Signatures of gate-tunable superconductivity in trilayer graphene moiré superlattice. *Nature* **572**, 215–219 (2019).
- Kopnin, N., Heikkilä, T. & Volovik, G. High-temperature surface superconductivity in topological flat-band systems. *Phys. Rev. B* **83**, 220503 (2011).
- Sharpe, A. L. et al. Emergent ferromagnetism near three-quarters filling in twisted bilayer graphene. *Science* **365**, 605–608 (2018).
- Xie, M. & MacDonald, A. H. On the nature of the correlated insulator states in twisted bilayer graphene. Preprint at <https://arxiv.org/abs/1812.04213> (2018).
- Ochi, M., Koshino, M. & Kuroki, K. Possible correlated insulating states in magic-angle twisted bilayer graphene under strongly competing interactions. *Phys. Rev. B* **98**, 081102 (2018).
- Dodaro, J. F., Kivelson, S. A., Schattner, Y., Sun, X.-Q. & Wang, C. Phases of a phenomenological model of twisted bilayer graphene. *Phys. Rev. B* **98**, 075154 (2018).
- Thomson, A., Chatterjee, S., Sachdev, S. & Scheurer, M. S. Triangular antiferromagnetism on the honeycomb lattice of twisted bilayer graphene. *Phys. Rev. B* **98**, 075109 (2018).
- Nandkishore, R., Levitov, L. & Chubukov, A. Chiral superconductivity from repulsive interactions in doped graphene. *Nat. Phys.* **8**, 158 (2012).
- Cao, Y. et al. Strange metal in magic-angle graphene with near Planckian dissipation. Preprint at <https://arxiv.org/abs/1901.03710> (2019).
- Po, H. C., Zou, L., Senthil, T. & Vishwanath, A. Faithful tight-binding models and fragile topology of magic-angle bilayer graphene. *Phys. Rev. B* **99**, 195455 (2019).
- Kim, K. et al. Tunable moiré bands and strong correlations in small-twist-angle bilayer graphene. *Proc. Natl Acad. Sci. USA* **114**, 3364–3369 (2017).
- Yankowitz, M. et al. Tuning superconductivity in twisted bilayer graphene. *Science* **363**, 1059–1064 (2019).
- Cao, Y. et al. Superlattice-induced insulating states and valley-protected orbits in twisted bilayer graphene. *Phys. Rev. Lett.* **117**, 116804 (2016).
- Lian, B., Xie, F. & Bernevig, B. A. The Landau level of fragile topology. Preprint at <https://arxiv.org/abs/1811.11786> (2018).
- Song, Z. et al. All magic angles in twisted bilayer graphene are topological. *Phys. Rev. Lett.* **123**, 036401 (2019).
- Bultinck, N., Chatterjee, S. & Zaletel, M. P. Anomalous Hall ferromagnetism in twisted bilayer graphene. Preprint at <https://arxiv.org/abs/1901.08110> (2019).
- Zhang, Y.-H., Mao, D. & Senthil, T. Twisted bilayer graphene aligned with hexagonal boron nitride: anomalous Hall effect and a lattice model. Preprint at <https://arxiv.org/abs/1901.08209> (2019).
- Polshyn, H. et al. Phonon scattering dominated electron transport in twisted bilayer graphene. Preprint at <https://arxiv.org/abs/1902.00763> (2019).
- Kondo, T. et al. Quadratic Fermi node in a 3D strongly correlated semimetal. *Nat. Commun.* **6**, 10042 (2015).
- Kagawa, F., Itou, T., Miyagawa, K. & Kanoda, K. Magnetic-field-induced Mott transition in a quasi-two-dimensional organic conductor. *Phys. Rev. Lett.* **93**, 127001 (2004).
- Lian, B., Wang, Z. & Bernevig, B. A. Twisted bilayer graphene: a phonon driven superconductor. *Phys. Rev. Lett.* **122**, 257002 (2019).
- Wu, F., Hwang, E. & Sarma, S. D. Phonon-induced giant linear-in- $T$  resistivity in magic angle twisted bilayer graphene: ordinary strangeness and exotic superconductivity. *Phys. Rev. B* **99**, 165112 (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### Device fabrication

Extended Data Fig. 1 displays a step-by-step stacking process for the fabrication of twisted bilayer graphene (tBLG) with a graphite bottom gate. The hBN/tBLG/hBN/graphite stacks were exfoliated and assembled using a van der Waals assembly technique. Monolayer graphene, thin graphite and hBN flakes (around 10-nm thick) were first exfoliated on SiO<sub>2</sub> (around 300 nm)/Si substrate, followed by the ‘tear and stack’ technique<sup>29</sup> with a polycarbonate (PC)/polydimethylsiloxane (PDMS) stamp to obtain the final hBN/tBLG/hBN/graphite stack. The separated graphene pieces were rotated manually by a twist angle of around 1.2–1.3°. We purposefully chose a larger twist angle during the assembly of the heterostructure owing to the high risk of relaxation of the twist angle to random lower values. To increase the structural homogeneity, we further carried out a mechanical cleaning process to squeeze the trapped blister out and release the local strain<sup>30</sup> (Extended Data Fig. 2). To avoid the uncertainty induced by thermal expansion of the transfer stage, all the stacking process were carried out at a fixed temperature of 100 °C, except that the final stacks were released at 180 °C (the melting point of polycarbonate). We did not perform subsequent high-temperature annealing to avoid relaxation of the twist angle. We further patterned the stacks with PMMA resist and CHF<sub>3</sub> + O<sub>2</sub> plasma and exposed the edges of graphene, which was subsequently contacted by Cr/Au (5/50 nm) metal leads using electron-beam evaporation (Cr) and thermal evaporation (Au).

### Measurement

Transport measurements were carried out in a dilution refrigerator with a base temperature of 16 mK and a perpendicular magnetic field of up to 5 T. The dilution refrigerator was well filtered to avoid heating of the electrons in our devices. We use superconducting-type coaxial cables (around 2 m long; Lakeshore) from the room-temperature plate to the mixing chamber plate of the cryostat. We add on each line a pi filter (RS 239-191) at room temperature, and a powder filter (Leiden Cryogenics) as well as a two-stage resistor–capacitor filter on a printed circuit board ( $R = 1\text{ k}\Omega$ ,  $C = 100\text{ nF}$ ) at the mixing chamber plate. The total resistance of each line is about 2 k $\Omega$ . The sample is located in a copper box with coaxial feedthroughs.

Standard low-frequency lock-in techniques were used to measure the resistance  $R_{xx}$  and  $R_{xy}$  with an excitation current of about 1 nA at a frequency of 19.111 Hz. In the measurement of differential resistance  $dV/dI$ , an a.c. excitation current (around 0.5 nA) was applied through an a.c. signal (0.5 V) generated by the lock-in amplifier in combination with a 1/100 divider and a 10-M $\Omega$  resistor. Before combining with the excitation, the applied d.c. signal passed through a 1/100 divider and a 1-M $\Omega$  resistor. As-induced differential voltage was further measured at the same frequency of 19.111 Hz with standard lock-in technique. For measurements in strong magnetic fields we found that the increased contact resistance made it difficult to obtain accurate values of the

device resistance. To resolve this issue, we applied a global gate voltage (+20 V) through Si/SiO<sub>2</sub> (around 300 nm) to tune the charge carrier density separately in the device leads.

### Twist angle extraction

The total carrier density  $n$  tuned by gate is calibrated by Hall measurements at low field (Extended Data Fig. 3). Near charge neutrality and band insulating states, Hall charge carrier density ( $n_H = -B/(eR_{xy})$ ) should closely follow gate-induced carrier density  $n$ ; that is,  $dn_H/dn = 1$ , providing accurate measurements of the carrier density  $n$ .

For different integer ( $\nu$ ) moiré filling states, the total carrier density can be described by  $\nu n_0 = \nu A_0^{-1} = 4\nu(1 - \cos\theta)/\sqrt{3}a^2$ , where  $A_0$  is the unit cell area of the periodic moiré pattern,  $\theta$  is the twist angle and  $a = 0.246\text{ nm}$  is the lattice constant of graphene. The local twist angles between different contacts are extracted with the carrier densities of  $\nu = 2$  states shown in Extended Data Fig. 4. The carrier density difference between CNP and  $\nu = 2$  states in device D1 ranges from  $1.38 \times 10^{12}\text{ cm}^{-2}$  to  $1.45 \times 10^{12}\text{ cm}^{-2}$ , corresponding to local twist angles ranging from 1.09° to 1.12°. For device D2, the local twist angles range from 1.08° to 1.10°.

### Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

29. Kim, K. et al. van der Waals heterostructures with high accuracy rotational alignment. *Nano Lett.* **16**, 1989–1995 (2016).
30. Purdie, D. G. et al. Cleaning interfaces in layered materials heterostructures. *Nat. Commun.* **9**, 5387 (2018).

**Acknowledgements** We are grateful for discussions with P. Jarillo-Herrero, A. Bernevig, M. Yankowitz, A. Young, C. Dean, L. Levitov, A. Vishwanath, M. Fisher, M. Allan and F. Koppens. D.K.E. acknowledges support from the Ministry of Economy and Competitiveness of Spain through the ‘Severo Ochoa’ program for Centres of Excellence in R&D (SE5-0522), Fundació Privada Cellex, Fundació Privada Mir-Puig, the Generalitat de Catalunya through the CERCA program, the H2020 Programme under grant agreement number 820378, Project: 2D-SIPC and the La Caixa Foundation. A.H.M. and M.X. acknowledge support from Department of Energy grant DE-FG02-02ER45958 and Welch Foundation grant TBF1473. A.B. acknowledges support from the Plan Nacional (RTI2018-097953-B-I00) of MICINN. G.Z. acknowledges support from the National Science Foundation of China under grant numbers 11834017 and 61888102, and the Strategic Priority Research Program of the Chinese Academy of Sciences under grant number XDB30000000.

**Author contributions** D.K.E. and X.L. conceived and designed the experiments; X.L., W.Y. and P.S. performed the experiments; X.L. and D.K.E. analysed the data; M.X. and A.H.M. performed the theoretical modelling of the data; T.T. and K.W. contributed materials; D.K.E., A.B., M.A.A., I.D., C.U. and G.Z. supported the experiments; X.L., D.K.E., P.S., X.M. and A.H.M. wrote the paper.

**Competing interests** The authors declare no competing interests.

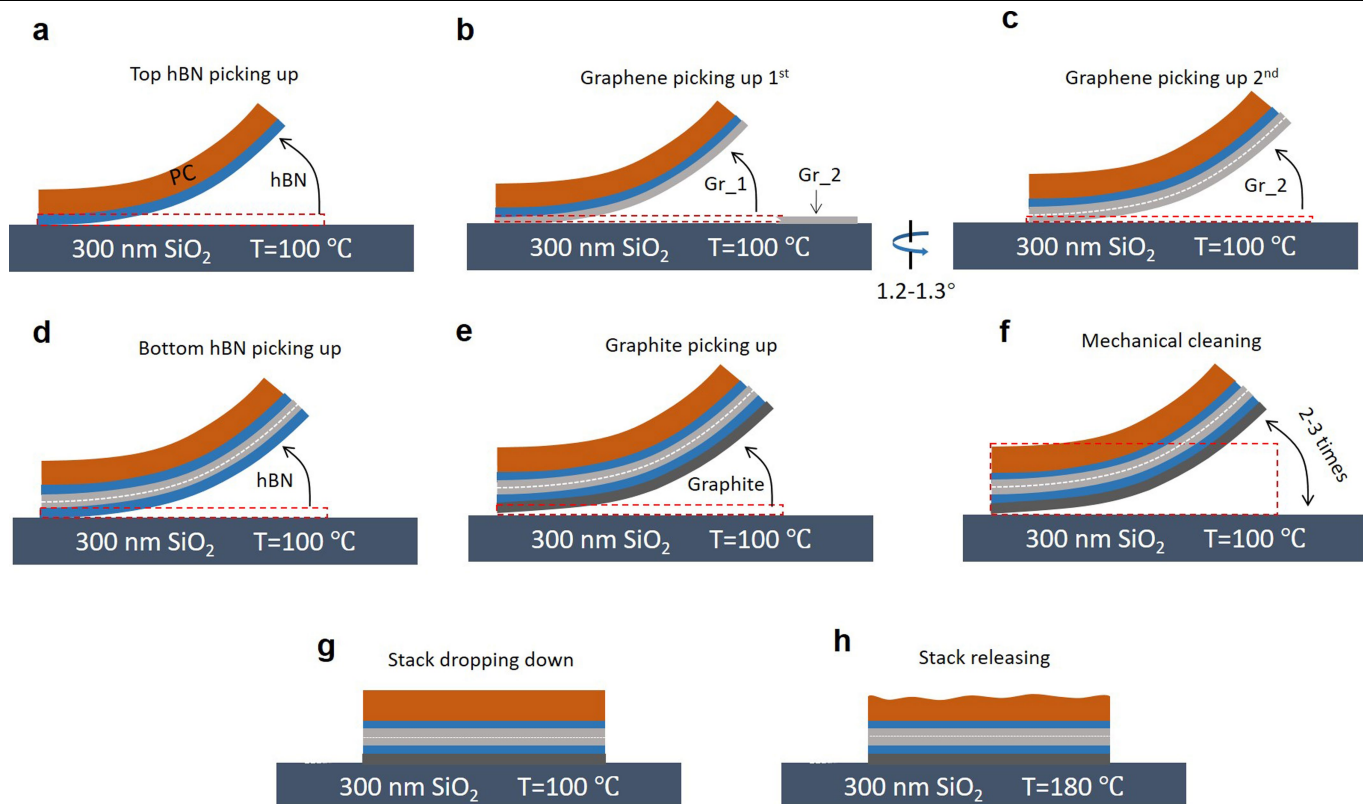
### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1695-0>.

**Correspondence and requests for materials** should be addressed to D.K.E.

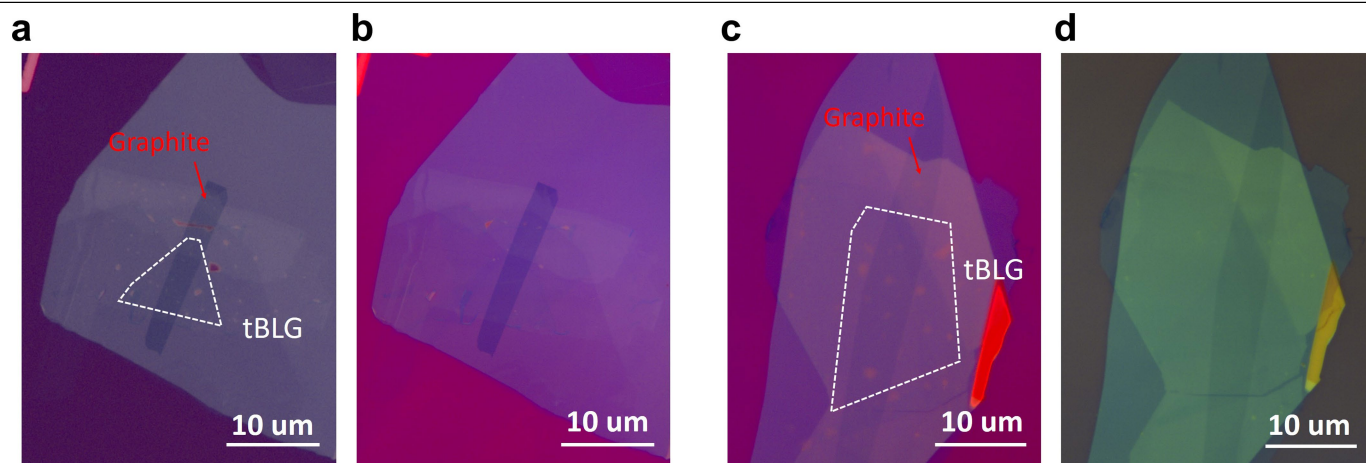
**Peer review information** Nature thanks David Goldhaber-Gordon and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

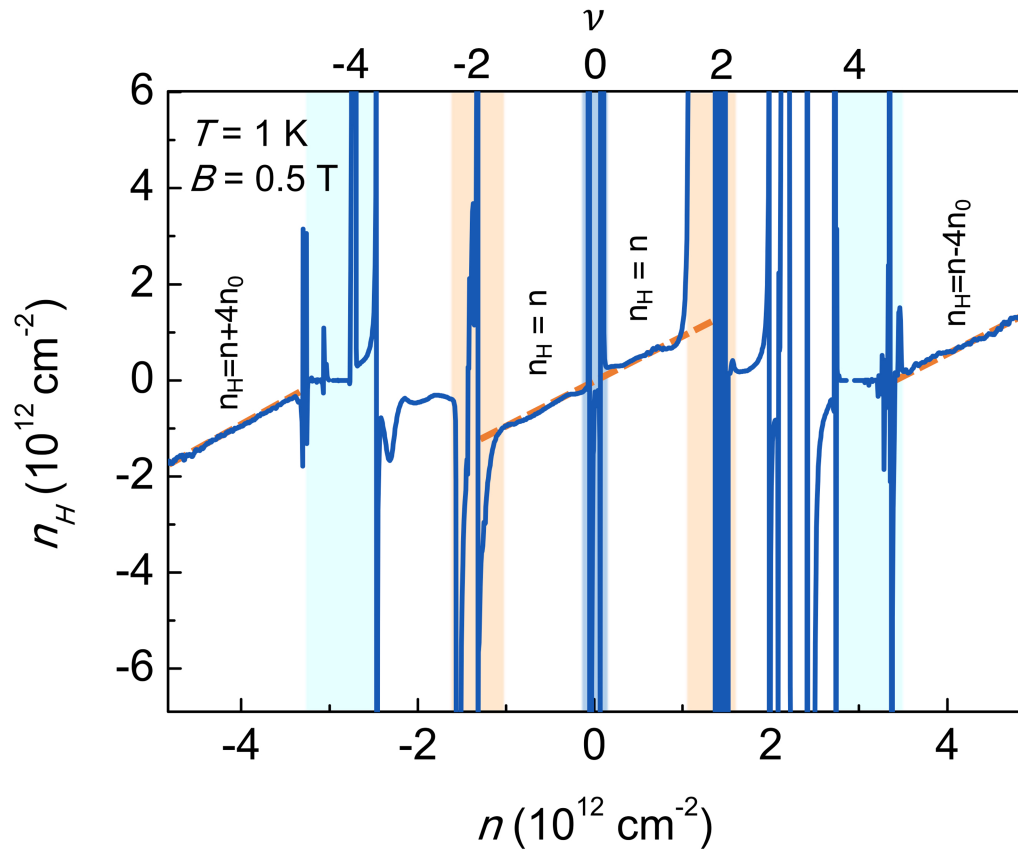


**Extended Data Fig. 1 | Schematic of the stacking process for the fabrication of twisted bilayer graphene with graphite bottom gate. a–h,** Sequential device fabrication method, describing the tear-and-stack co-lamination process used to create the hBN/tBLG/hBN/graphite stacks.



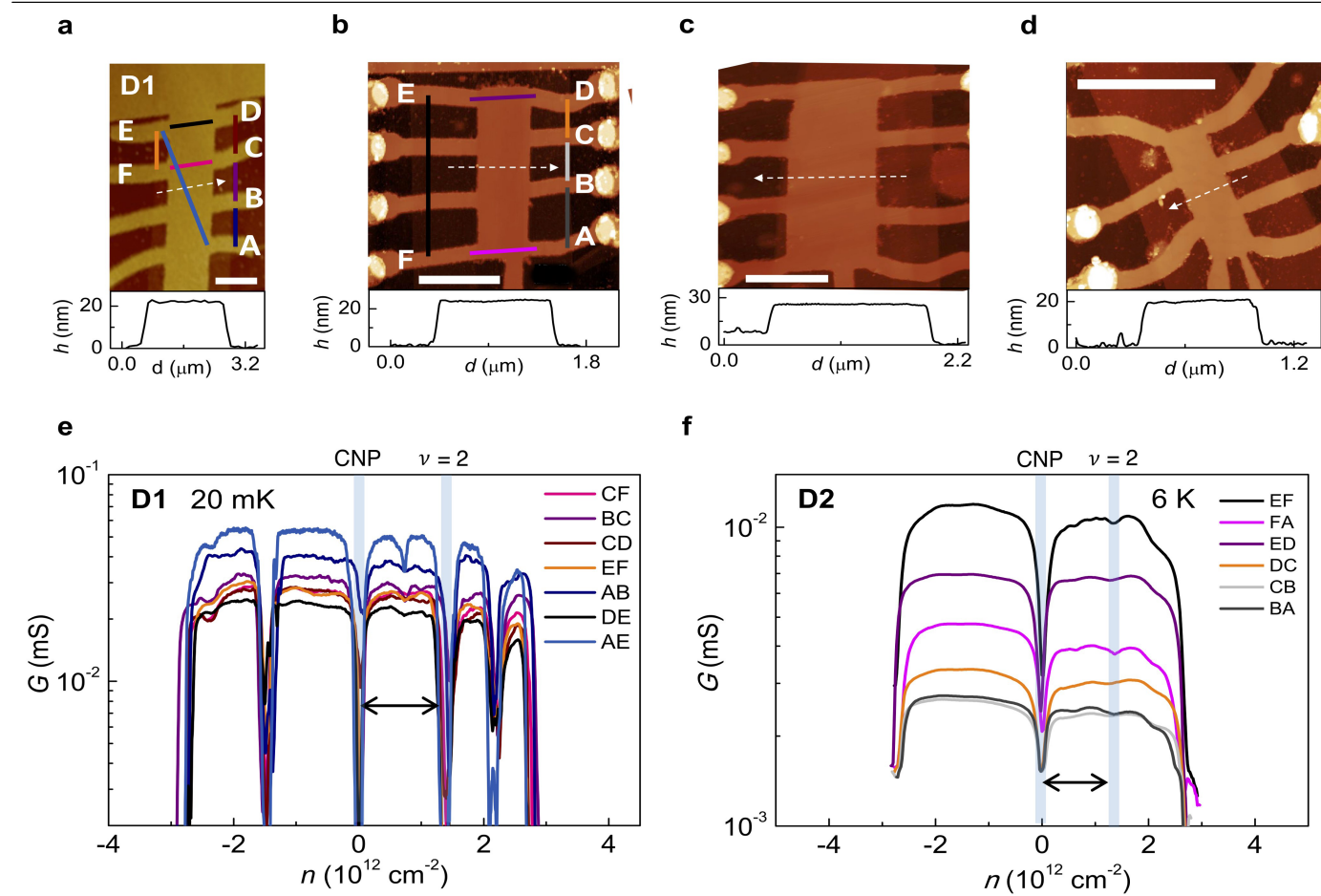


**Extended Data Fig. 2 | Mechanical cleaning of twisted bilayer graphene. a–d,** Optical images of the final stacks before mechanical cleaning (a, c) and after mechanical cleaning (b, d).



**Extended Data Fig. 3 | Hall measurements of device D1.** Coloured vertical bars correspond to filling factors  $\nu = -4, -2, 2$  and  $4$ . Hall charge carrier density ( $n_H = -B/(eR_{xy})$ ) closely follows the gate-induced carrier density  $n$ . Near charge

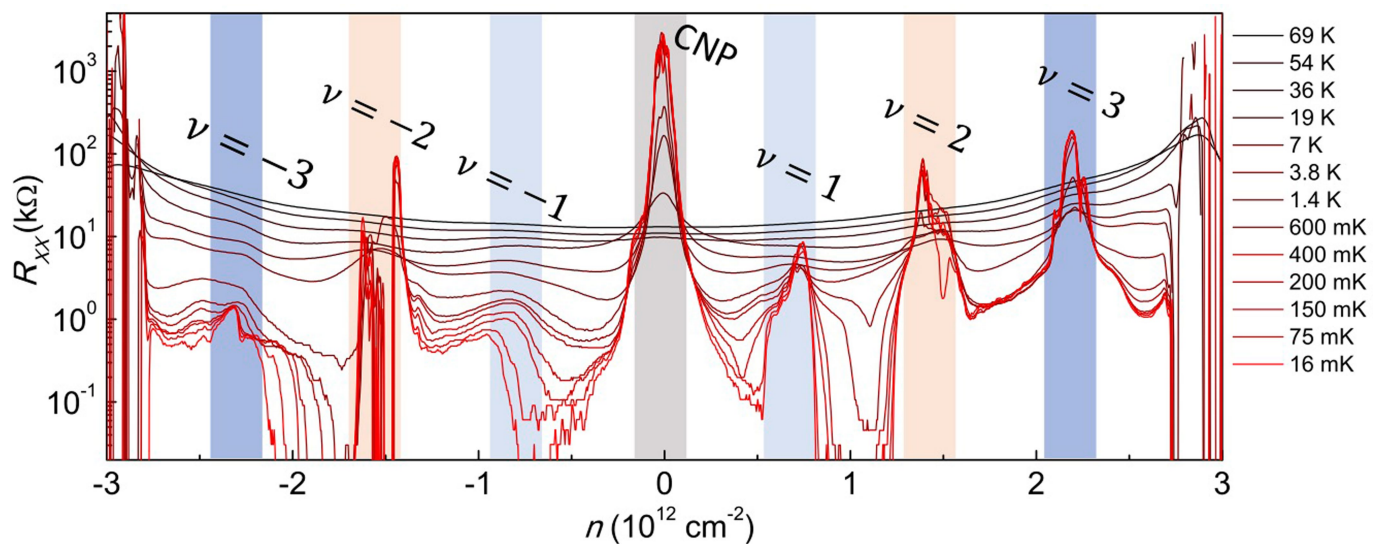
neutrality,  $n_H = n$ . Beyond the band insulator regions ( $\nu = \pm 4$ ), the Hall density strictly follows  $n_H = n \pm 4n_0$ .



**Extended Data Fig. 4 | Measuring the homogeneity of the twist angle.**  
**a–d**, Atomic force microscopy images of a set of twisted bilayer graphene samples. Scale bar, 2  $\mu\text{m}$ . Dashed-line arrows correspond to the height profiles shown below the topographies. **e, f**, Two-terminal conductance measurements

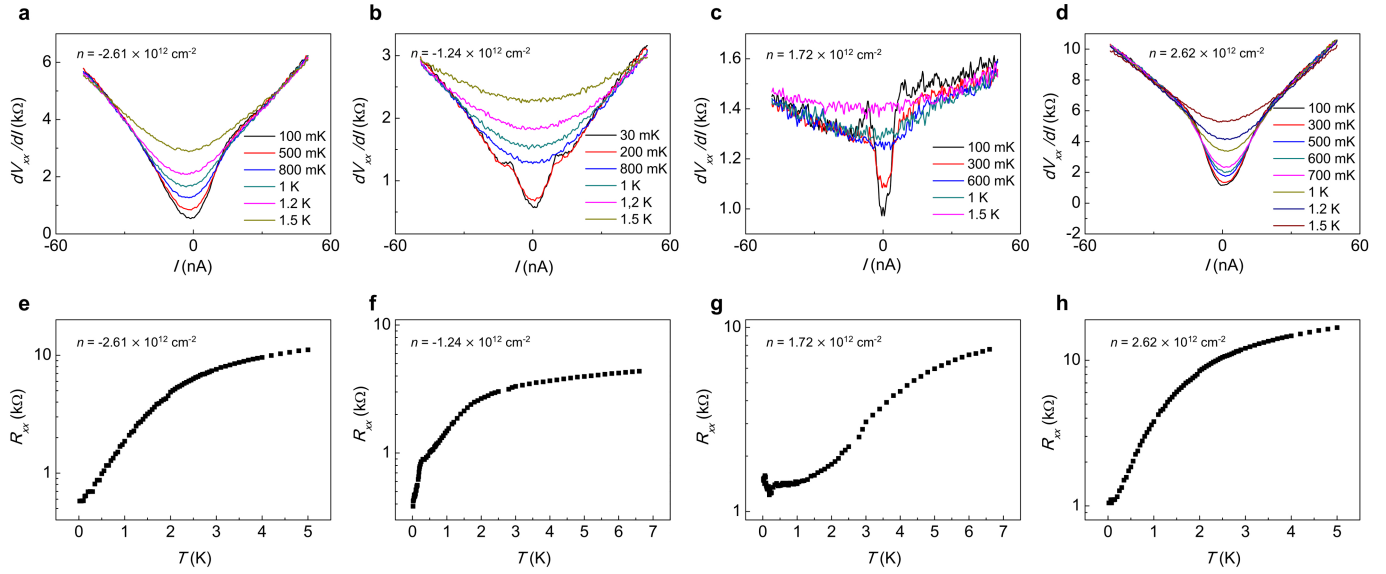
taken between contacts shown in **a** and **b**. Colours correspond to the bars shown in **a** and **b**, respectively. The difference in carrier density between the CNP and the  $\nu = 2$  state is used to extract the local twist angle.





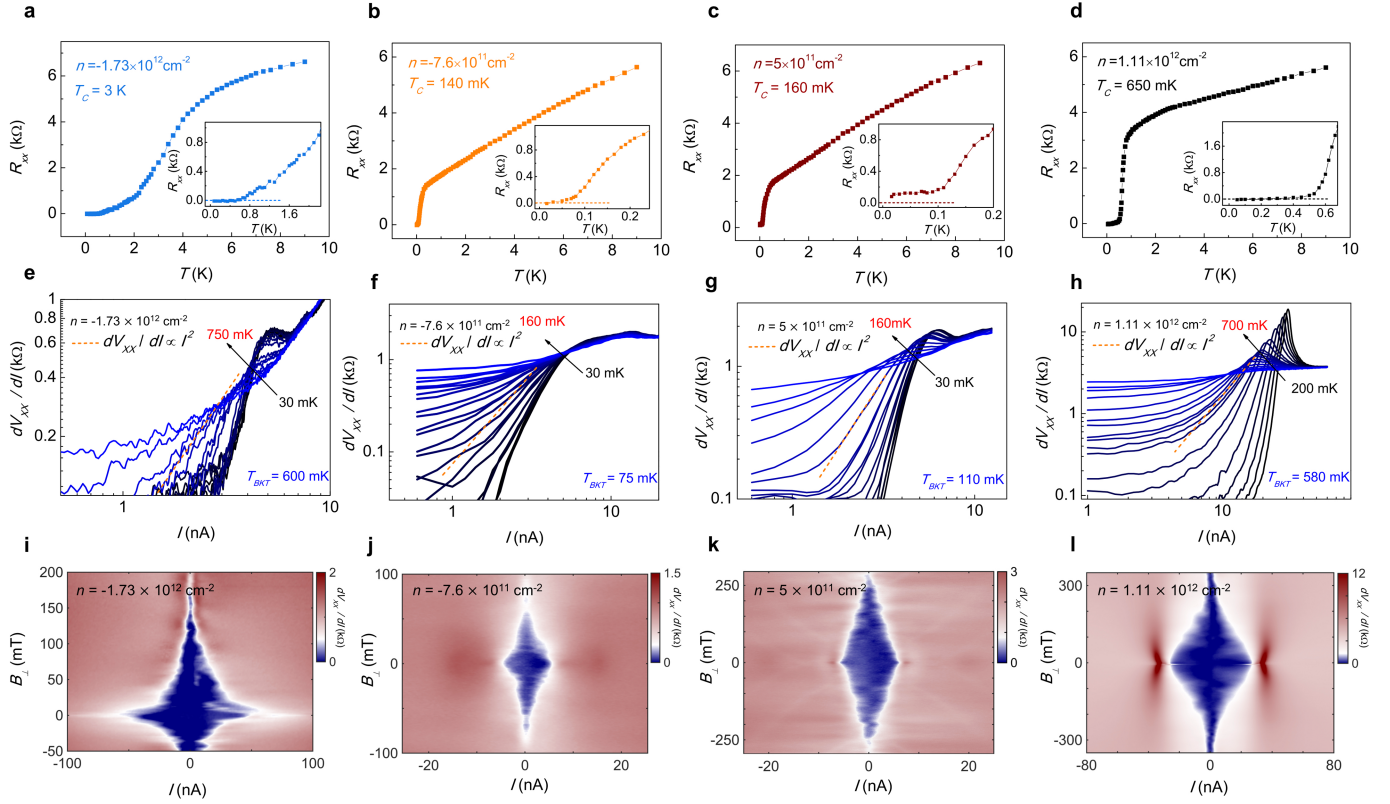
**Extended Data Fig. 5 | Four-terminal longitudinal resistance as a function of carrier density at different temperatures.** The four-terminal longitudinal resistance is plotted against carrier density  $n$  for different temperatures, from

69 K (black trace) to 16 mK (red trace). Coloured vertical bars correspond to the filling factors  $\nu$  as shown.



**Extended Data Fig. 6 | Additional measurements of other possible superconducting domes. a–d**, Differential resistance measurements for additional domes between  $-4n_0$  and  $-3n_0$  (a),  $-2n_0$  and  $-n_0$  (b)  $2n_0$  and  $3n_0$  (c) and

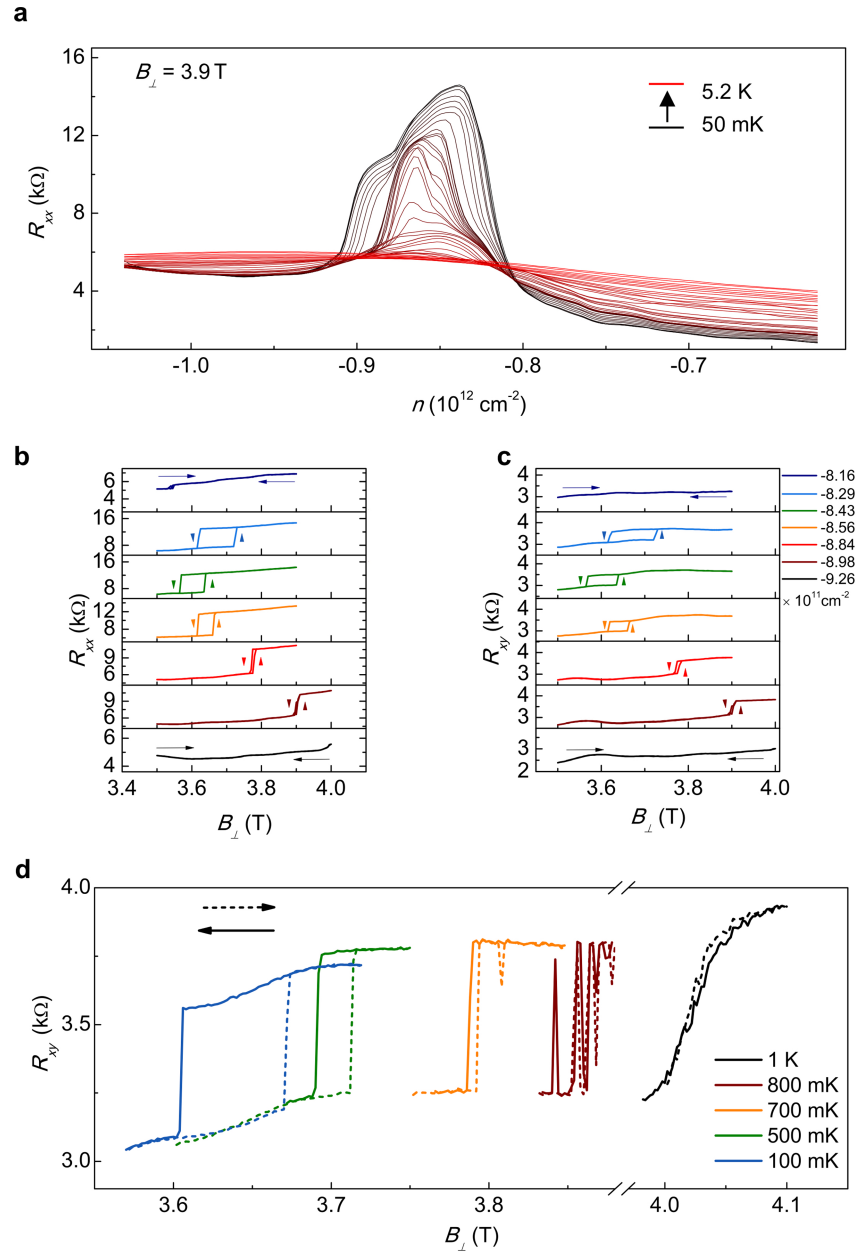
$3n_0$  and  $4n_0$  (d). **e–h**, Corresponding thermal activation measurements of resistance against temperature for the same carrier densities as in **a–d**, respectively.



**Extended Data Fig. 7 | Full characterization of all four superconducting pockets in sample D1.** **a–d**, Thermal activation measurements of resistance against carrier density. The inset shows magnified images, demonstrating that in three superconducting states the resistance drops completely to zero (**a**, **b**, **d**) and in one superconducting state the resistance saturates at about  $80 \Omega$  (**c**).

**e–h**, Differential resistance is plotted against d.c. bias current at various temperatures in order to establish BKT transition temperatures. **i–l**, Two-dimensional colour plots of the differential resistance as a function of magnetic field and excitation current at 16 mK.





**Extended Data Fig. 8 | Additional magnetic hysteresis data. a**, Four-terminal longitudinal resistance as a function of carrier density at different temperatures from 50 mK (black trace) to 5.2 K (red trace). **b, c**, Plots of longitudinal resistance  $R_{xx}$  and transverse resistance  $R_{xy}$  against magnetic field at different charge carrier densities and 100 mK. Arrows indicate the sweep direction of the

magnetic field. Data from **b** is used to extract data for Fig. 4e. **d**, Transverse resistance plotted against magnetic field at different temperatures (the same dataset as in Fig. 4c). Dashed and solid lines correspond to ascending and descending magnetic fields, respectively.

Extended Data Table 1 | Full dataset for all observed superconducting states in device D1

SC pocket density	$T_c$ (mK)	$T_{BKT}$ (mK)	$B_{C(T=0)}$ (mT)	$\xi_{GL(T=0)}$ (nm)
$-1.73 \times 10^{12} \text{ cm}^{-2}$	3000	600	$\sim 180$	$\sim 41$
$-7.6 \times 10^{11} \text{ cm}^{-2}$	140	75	$\sim 100$	$\sim 55$
$5 \times 10^{11} \text{ cm}^{-2}$	160	110	$\sim 300$	$\sim 32$
$1.11 \times 10^{12} \text{ cm}^{-2}$	650	580	$\sim 400$	$\sim 27$

# Polymers with controlled assembly and rigidity made with click-functional peptide bundles

<https://doi.org/10.1038/s41586-019-1683-4>

Received: 19 April 2018

Accepted: 14 August 2019

Published online: 30 October 2019

Dongdong Wu<sup>1</sup>, Nairiti Sinha<sup>1</sup>, Jeeyoung Lee<sup>1</sup>, Bryan P. Sutherland<sup>1</sup>, Nicole I. Halaszynski<sup>1</sup>, Yu Tian<sup>1</sup>, Jeffrey Caplan<sup>2</sup>, Huixi Violet Zhang<sup>3</sup>, Jeffery G. Saven<sup>3\*</sup>, Christopher J. Kloxin<sup>1,4\*</sup> & Darrin J. Pochan<sup>1\*</sup>

The engineering of biological molecules is a key concept in the design of highly functional, sophisticated soft materials. Biomolecules exhibit a wide range of functions and structures, including chemical recognition (of enzyme substrates or adhesive ligands<sup>1</sup>, for instance), exquisite nanostructures (composed of peptides<sup>2</sup>, proteins<sup>3</sup> or nucleic acids<sup>4</sup>), and unusual mechanical properties (such as silk-like strength<sup>3</sup>, stiffness<sup>5</sup>, viscoelasticity<sup>6</sup> and resiliency<sup>7</sup>). Here we combine the computational design of physical (noncovalent) interactions with pathway-dependent, hierarchical ‘click’ covalent assembly to produce hybrid synthetic peptide-based polymers. The nanometre-scale monomeric units of these polymers are homotetrameric,  $\alpha$ -helical bundles of low-molecular-weight peptides. These bundled monomers, or ‘bundlemers’, can be designed to provide complete control of the stability, size and spatial display of chemical functionalities. The protein-like structure of the bundle allows precise positioning of covalent linkages between the ends of distinct bundlemers, resulting in polymers with interesting and controllable physical characteristics, such as rigid rods, semiflexible or kinked chains, and thermally responsive hydrogel networks. Chain stiffness can be controlled by varying only the linkage. Furthermore, by controlling the amino acid sequence along the bundlemer periphery, we use specific amino acid side chains, including non-natural ‘click’ chemistry functionalities, to conjugate moieties into a desired pattern, enabling the creation of a wide variety of hybrid nanomaterials.

Our bundlemer-based polymer chains exhibit a variety of unique features. Unlike high-molecular-weight synthetic polymers, our chains use small (roughly 3 kDa), easily synthesized peptide sequences that fold into designed tetrameric 4-nanometre bundles. The subsequent covalent assembly of these bundles yields polymers with micrometre-scale contour lengths. The design of  $\alpha$ -helical homo-oligomers has a long history, with both empirical *de novo*<sup>8</sup> and computational<sup>9</sup> methods being used. Here, computationally designed homotetrameric bundles with D<sub>2</sub> symmetry<sup>10</sup> present two reactive groups at each end, owing to chemical functionalization of the amino termini of the constituent peptides (Fig. 1a). Distinct homotetrameric bundles with complementary reactive functional groups are chemically linked (or ‘clicked’ together) to produce bundlemer chains.

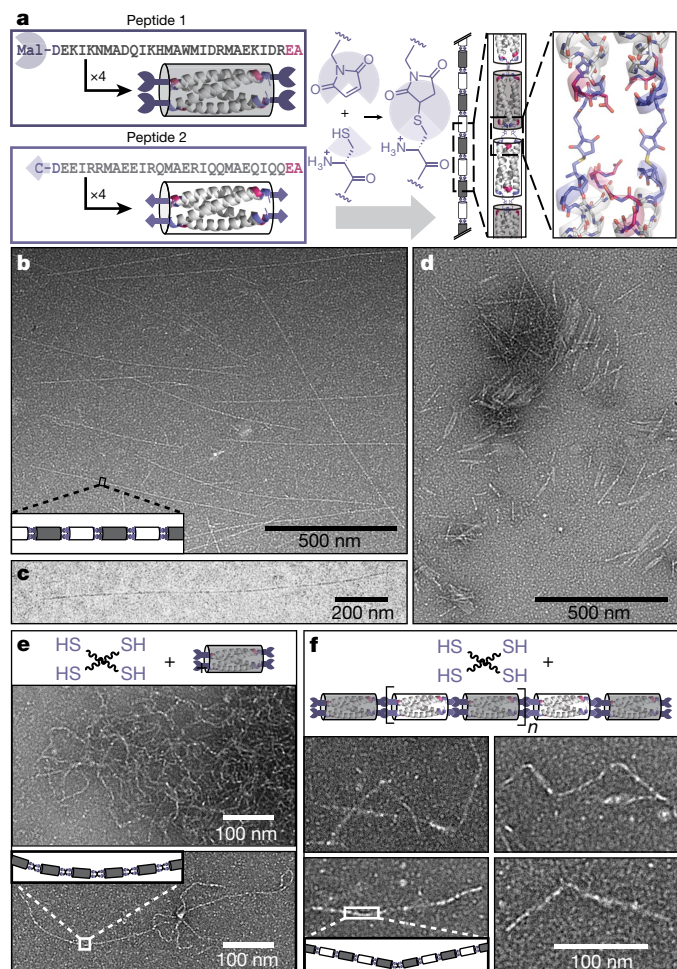
Our first example of a bundlemer chain uses two different coiled-coil peptide bundles<sup>10</sup> with complementary reactive groups for covalent polymerization (Fig. 1a). One peptide (peptide 1; Extended Data Fig. 1) contains the non-natural organic group maleimide at the N terminus, resulting in two maleimides at each end of the folded bundlemer. The second peptide contains the amino acid cysteine at the N terminus

(peptide 2; Extended Data Fig. 1). The maleimide and the thiol group of cysteine undergo a thiol-Michael addition reaction (Extended Data Fig. 2) to form two covalent bonds between neighbouring bundlemers, producing a hybrid polymer chain that is maintained by both the covalent linkages and the complementary, noncovalent (physical) interactions within each bundlemer. The resulting polymers exhibit an unprecedented rigid-rod character. Both negatively stained transmission electron microscopy (TEM) and cryogenic TEM (cryoTEM) data reveal extreme chain stiffness (Fig. 1b–d). Direct imaging of the rigid rods is possible because the polymer cross-section is defined by constituent bundlemers. Small-angle neutron scattering (SANS) of the rigid-rod polymers confirms that the rod cross-section is that of a single bundlemer with a diameter of roughly 2 nm (Extended Data Fig. 3).

Although one-dimensional (1D) assemblies of proteins can exhibit a high rigidity (as in viruses<sup>11,12</sup>, actin<sup>13,14</sup>, microtubulin<sup>13</sup> and misfolded globular proteins<sup>15</sup>), all are assemblies of much larger proteins, with many interprotein interactions, much wider cross-sections, and large ratios of protein molecular mass to linear length of the rod<sup>16</sup>. Similarly, other physically assembled coiled-coils<sup>17–19</sup> have produced stiff fibres

<sup>1</sup>Department of Materials Science and Engineering, University of Delaware, Newark, DE, USA. <sup>2</sup>Delaware Biotechnology Institute, University of Delaware, Newark, DE, USA. <sup>3</sup>Department of Chemistry, University of Pennsylvania, Philadelphia, PA, USA. <sup>4</sup>Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, DE, USA. \*e-mail: saven@upenn.edu; cjk@udel.edu; poch@udel.edu





**Fig. 1 | Peptides designed for homotetrameric, antiparallel coiled-coil bundle formation<sup>25</sup> that form polymers upon covalent assembly.** **a**, Left, peptides 1 and 2 (Extended Data Fig. 1), shown in single-letter amino acid code, have at their N termini (blue) either maleimide (Mal) or cysteine (C). The carboxyl terminus (red) of each peptide is unreactive. Each sequence forms homotetrameric bundlemers: grey, peptide 1; white, peptide 2. Centre, the thiol–maleimide click reaction yields chains with two covalent linkages between neighbouring bundlemers. **b**, TEM of rigid rods produced with a 1/1 ratio of peptides 1 and 2. The sample is negatively stained with phosphotungstic acid (PTA). **c**, CryoTEM of rigid rods longer than 1  $\mu\text{m}$  in aqueous solution. **d**, Negatively stained TEM of short rigid-rod chains produced using an asymmetric ratio (10/9: [peptide 1]/[peptide 2]) of reacting bundlemers. **e**, The organic tetrathiol PETMP (black wavy lines) links peptide-1 bundlemers to form semiflexible chains. **f**, Examples of segmented chains produced by connecting short rigid rods with PETMP. Rod segments within the segmented polymers range in length from approximately 50 nm (where  $n$ , the number of bundlemers per segment, is approximately 3 to 4) to 100 nm (where  $n$  is approximately 8 to 9).

with large cross-sections relative to the protein or peptide building blocks. Natural fibrils that are rich in  $\beta$ -sheets made from misfolded proteins—such as bovine serum albumin<sup>20,21</sup> or  $\beta$ -lactoglobulin<sup>22</sup>—as well as short de novo designed peptides<sup>23</sup> exhibit large persistence lengths ( $l_p$ ) that range from 10 nm for thinner fibrils up to approximately 10  $\mu\text{m}$  for thicker fibrils. However, the stiffer, thicker  $\beta$ -sheet fibrils must have a substantial mass per length in order to produce such a high- $l_p$  1D nanostructure. The biopolymer with the cross-section that is closest to that observed here is double-stranded DNA; however, DNA has an  $l_p$  of 50 nm or less (depending on solution conditions)<sup>24</sup>. The TEM and cryoTEM data in Fig. 1 reveal rods longer than 1  $\mu\text{m}$  that display rigid-rod behaviour along the entire chain length. Estimates of the rod persistence length that are based on methods developed for rod and fibre analysis in two

dimensions<sup>21</sup> provide values of more than 30  $\mu\text{m}$  (Extended Data Fig. 4), highlighting the extraordinary stiffness of such a thin molecular object. To put this in context, in Extended Data Fig. 5 we plot the persistence length versus the mass per unit length for our bundlemer rigid rods and for the other nanostructures cited above.

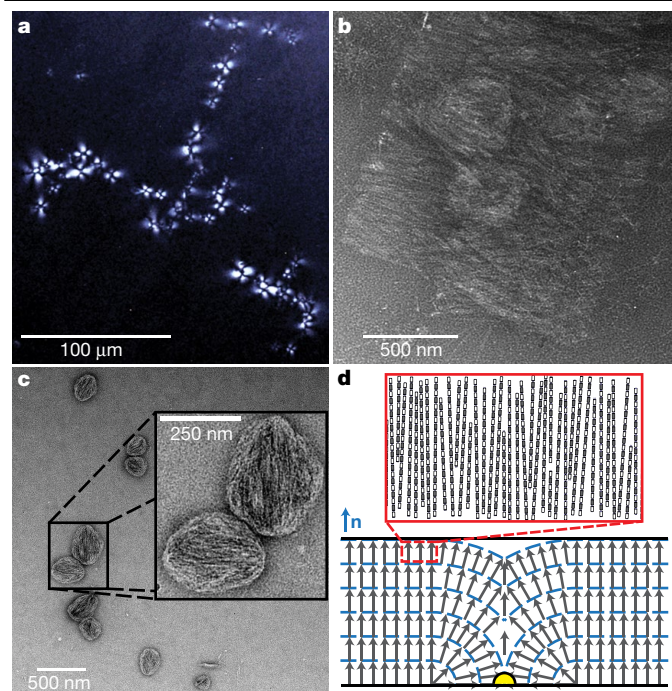
Hybrid physical–covalent interactions between proteins have until now produced flexible chains, owing to the flexibility of the linear polymer linkers<sup>25</sup> or dimeric coiled-coils<sup>26</sup> that link the final polymer together. Hexameric peptide coiled-coils have produced semiflexible nanotube chains through electrostatic interactions<sup>27</sup> or native chemical ligation; these semiflexible nanotubes exhibited an irregular chain trajectory and were shorter than the rigid rods produced here. An important aspect of our bundlemer chain rigidity is the use of antiparallel homotetrameric bundlemers as the building blocks for polymerization. Formation of both possible covalent linkages between two bundlemers is highly probable (Fig. 1a), because once one covalent bond between the bundlemers has formed, a second crosslink between the same pair is more likely than a linkage to a third bundle. The thiol and maleimide reaction sites are closely located at the ends of the rigid bundlemers, and the dual linkage between adjacent bundlemers restricts conformational flexibility within the final rigid-rod chain. This designed crosslinking contrasts with linkages among parallel hexameric coiled-coils, where defects in linking produce semiflexible chains with possible branch points<sup>27</sup>. The design of two covalent linkages between adjacent peptide bundles with  $D_2$  symmetry minimizes chain defects and provides extreme polymer rigidity.

The efficiency of the thiol–Michael reaction allows the targeting of rigid-rod length through the simple stoichiometric control of reactants. The ultralong (micrometre-size) rigid rods in Fig. 1b, c were synthesized with a 1/1 ratio of bundlemers displaying thiol (peptide 2) and maleimide (peptide 1) groups. By simply altering the bundle ratio (to 10/9; see Methods), much shorter rigid rods can be produced (Fig. 1d).

Just as we are not limited to a single bundlemer building block or polymer length, we also can use different organic interbundle linkers to alter the physical characteristics of the chain, while still producing a targeted bundlemer pattern along the chain. For example, bundles functionalized with maleimide (peptide 1) react with the organic tetrathiol pentaerythritol tetrakis(3-mercaptopropionate) (PETMP; Extended Data Fig. 6), to produce a semiflexible chain (Fig. 1e). The chain flexibility must be due solely to the molecular flexibility of the PETMP, because the same maleimide-functionalized bundlemer was also used to make rigid-rod chains. SANS of the semiflexible chains in solution indicates a chain cross-section of one peptide bundle, as observed with the rigid rods (Extended Data Fig. 3). We can also combine rigid-rod and semiflexible segments within a single polymer. Short rigid rods with maleimide-functionalized bundlemers at the rod termini can be reacted with PETMP to produce kinked, segmented chains (Fig. 1f), with the kinks resulting from the conformational flexibility of PETMP. Thus different polymer architectures and flexibilities are possible using the same bundlemer building blocks, thereby separating the characteristics of the chain contour from the designed amino acid sequence.

The hybrid physical–covalent peptide rigid rods exhibit lyotropic liquid crystalline behaviour in concentrated solution, with optical textures typical of lamellar phases. Toric focal conic domains (TFCDs) can be observed in smectic liquid crystals when confined to thin films in which the smectic layers are generally parallel with the sample substrate<sup>28</sup>. Here, we observe apparent TFCDs in polarizing optical microscopy of thin films of concentrated rod solutions (Fig. 2). Peptide or protein fibrillar assemblies are known to form liquid-crystal phases owing to their inherent rod-like character<sup>29</sup>. Given the extreme stiffness of our new physical–covalent bundlemer rods, the ability to target desired distributions of rod length, and the ability to alter rod chemistry using computational design and non-natural amino acids, these systems provide new opportunities for liquid-crystal material design.

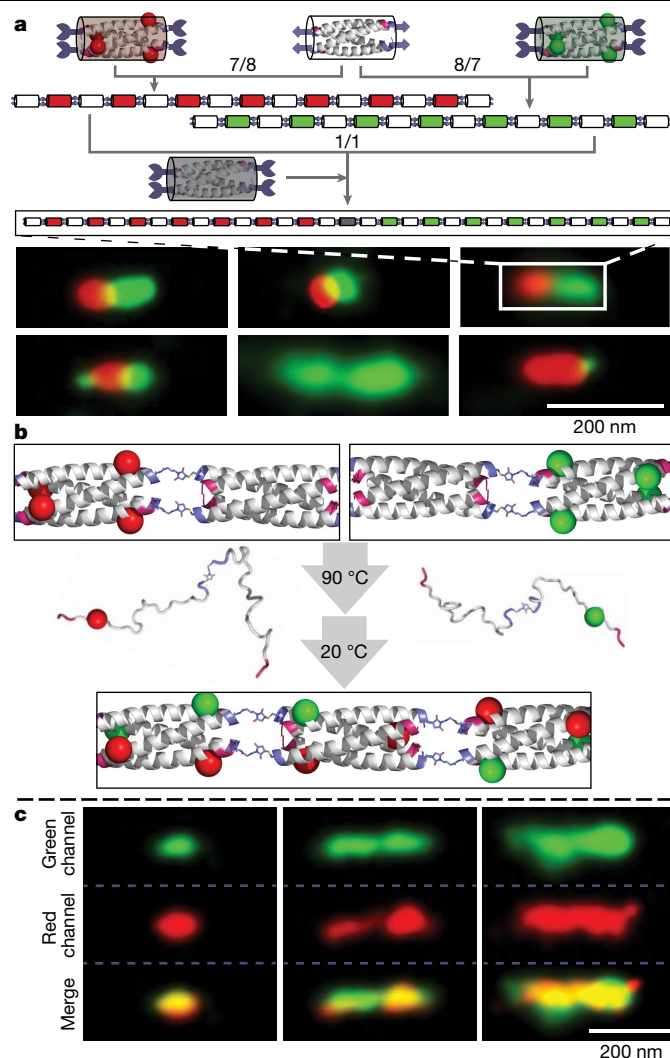
We can purposefully display and pattern chemical moieties ranging from small-molecule dyes to synthetic polymers to inorganic



**Fig. 2 | Liquid-crystal behaviour of concentrated rigid-rod solutions.** Rods were prepared as in Fig. 1a, d with alternating bundlemers of peptide 1 and peptide 2. **a**, Polarized optical microscopy (POM) of a pseudoisotropic region of roughly 100-nm-long rods with multiple TFCs, indicative of a lyotropic lamellar phase. POM was performed on an 8% (w/v) short rigid-rod solution at pH 2. **b**, TEM of negatively stained short rigid rods from the pseudoisotropic region shown in **a**, revealing clear rod layering. **c**, TEM reveals the structure of dilute regions in which rigid rods have locally aggregated into droplets with clear rod orientation. **d**, Bottom, diagram of a TFCD cross-section formed in smectic-A-type liquid crystals. Top, enlargement of a single smectic layer, showing the proposed homeotropic alignment of individual rigid rods. The blue dashed lines represent boundaries between smectic layers confined between parallel walls (thick black lines represent the glass slide and cover slip in the POM). The liquid-crystal director  $\mathbf{n}$ , the axis along which all rods are aligned within individual layers, is perpendicular to the smectic layers. The local orientation director (grey arrows) within the smectic A layers is parallel to  $\mathbf{n}$  far from the TFCD. In the vicinity of a topological defect on the glass substrate (yellow), the local orientation field folds towards the defect.

nanoparticles. Super-resolution fluorescence microscopy<sup>30</sup> provides the tool for observing rigid rods synthesized with fluorophores attached to specific lysine side chains. We first assembled rods with targeted lengths of around 50 nm (Fig. 1), with the rod termini consisting of bundlemers displaying thiols (Fig. 3a; see Methods for more details). We then functionalized these short rods with either a green (peptide 3) or a red (peptide 4) fluorescent dye (green, 5(6)-carboxy-tetramethylrhodamine; red, 4-chloro-7-nitrobenzofurazan; Fig. 3a). We mixed separate populations of short red rods and short green rods with unlabelled bundlemers with maleimide termini (Figs. 1a, 3a) in order to link the short, dye-labelled rods into longer rigid rods. Stochastic optical reconstruction microscopy (STORM)<sup>30</sup> allows clear observation of the individual approximately 50-nm red or green segments within the longer rods (Fig. 3a).

Bundlemers within the rigid-rod chains can denature at high temperatures, resulting in disassembly of the chains, but the bundlemers can reversibly reform rigid rods below the bundlemer melting temperature. In the case of the rigid rods in Fig. 3, when the temperature was increased to 90 °C all constituent bundles denatured, causing the rigid rods to fall apart into covalently linked dimers of peptides, each dimer containing one dye-labelled peptide and one peptide without dye (Fig. 3b). When the temperature was reduced to 25 °C, the bundlemers reassembled

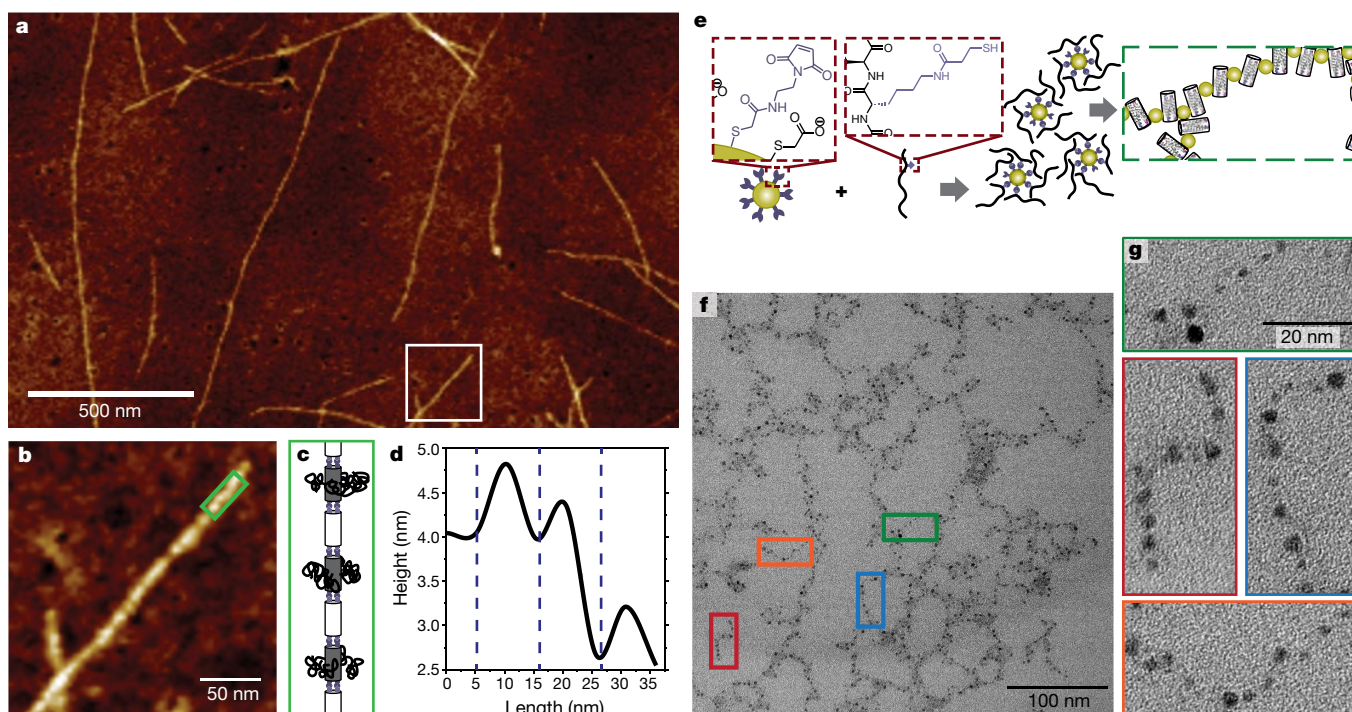


**Fig. 3 | Reversible noncovalent assembly of rod polymers.** **a**, Rigid rods were created using fluorescently labelled variants (peptide 3 (right) and peptide 4 (left); Extended Data Fig. 1), each containing either 4-chloro-7-nitrobenzofurazan (green) or 5(6)-carboxy-tetramethylrhodamine (red) attached to the lysine-24 side chain. Bundlemers of peptide 2 (centre, white) were used to form short rigid rods comprising a single dye type. The resulting red and green rods were joined with peptide-1 bundlemers (grey) to make longer rods with red and green segments. The STORM images below are of resulting longer rigid rods. The constituent red or green fluorescence of each segment is easily resolved. **b**, Rigid rods from **a** are heated to 90 °C, resulting in unfolding and dissociation of the individual bundles while peptide dimers remain covalently linked. When the solution is then cooled to 20 °C, the bundlemers and rigid rods reform. **c**, Reassembled rods now display co-localization of green and red fluorescence (producing a yellow signal when the green and red channels are displayed concurrently) along the entire reformed rod lengths.

and physically repolymerized into rigid rods. However, the dyes that were originally segregated into roughly 50-nm red or green segments within the longer rigid rods were now mixed along the entire length of the reformed rigid rods (Fig. 3c). This was a direct effect of the assembly pathway: the segregation of dyes within the rods created by the original pathway was erased by temperature changes to create a similar rigid-rod superstructure but with the dye segregation now scrambled.

The use of high-molecular-weight organic reagents to link bundlemers together<sup>6</sup> yields thermally responsive hydrogels. Maleimide-bearing bundlemers coupled to a 20-kDa tetrathiol with four poly(ethylene glycol) (PEG) arms (Extended Data Fig. 6) act as crosslinking points in a hybrid bundlemer-PEG hydrogel network. The type of bundlemer used





**Fig. 4 | Display of non-natural side chains by bundlers affords templated, patterned display of polymers from rigid rod bundlers or assembly of hybrid metal nanoparticle–bundle chains.** **a**, AFM image of rigid rods formed using peptides 2 and 6 (Extended Data Fig. 1), with azide-functionalized PEG<sub>2000</sub> chains conjugated to the rigid rods. **b**, AFM image of the rigid-rod area within the white outline in **a**; the area in the green rectangle was used for height analysis along the rod longitudinal axis (**d**). **c**, Diagram illustrating bundles of peptide 6

(grey) and peptide 2 (white) conjugated with PEG<sub>2000</sub>. **d**, Height trace along the longitudinal axis in **b**. **e**, Left, maleimide-functionalized gold nanoparticles are conjugated with peptide 7 (Extended Data Fig. 1), and then allowed to assemble into hybrid nanoparticle–bundlemer chains (right). **f**, TEM of nanoparticle–bundlemer chains. **g**, Magnified TEM images of the indicated nanoparticle–bundlemer chains in **f** reveal interparticle separation consistent with the dimensions of peptide bundles.

as the crosslinker controls the temperature response of the network. For example, the maleimide-terminated peptide-1 bundlemer—which has a melting temperature,  $T_m$ , of 55 °C (ref. <sup>10</sup>)—produces a hydrogel that can be reversibly disassembled and reformed by cycling the temperature above and below this temperature (Extended Data Fig. 7). By contrast, the peptide-5 bundlemer has a  $T_m$  of 80 °C (ref. <sup>10</sup>), and produces a network that is robust to much higher temperatures (Extended Data Fig. 7).

The structure of the bundlemer polymer can also be leveraged to pattern additional, orthogonal click reactions. Peptide 6 (Extended Data Fig. 1) contains a lysine at residue 13 that is functionalized to present an alkyne (Supplementary Information, scheme S5), which can react with an azide to form a triazole linkage via copper(I)-catalysed azide–alkyne cycloaddition (CuAAC). We further functionalized peptide-6-based bundlemer rods with azide-terminated PEG polymers (PEG<sub>2000</sub>). Atomic force microscopy (AFM) studies (Fig. 4) reveal height variation along the longitudinal rod axes, with a peak spacing of approximately 10 nm (Fig. 4d), consistent with the expected spacing of the PEG-functionalized bundles within the rigid rod.

Engineered click reactions can be used to link desired moieties to precise locations along existing bundlemer chains, but it is also possible to conjugate moieties to individual peptides before bundlemer formation, owing to the noncovalent self-assembly of the bundlers themselves. We designed peptide 7 (Extended Data Fig. 1) to display a thiol group only at residue 24 and conjugated it with maleimide-functionalized gold nanoparticles (Fig. 4). Subsequent aqueous assembly produces a dominant nanostructure of gold nanoparticle chains with an interparticle spacing of less than 5 nm, consistent with the interspacing of bundlers between adjacent nanoparticles. The hybrid nanoparticle–bundlemer chains exhibit variation in the apparent distances between the gold nanoparticles and changes in the chain trajectories

owing to the many possibilities for nanoparticle–bundlemer covalent connection.

The computational design of coiled-coil building blocks, combined with the engineering of covalent interactions, makes it possible to generate new types of physical–covalent programmable polymers, peptide liquid crystalline materials, and even one-, two- and three-dimensional nanostructures. The ‘bundlemer’ concept also allows the functionality of these structures to be altered with natural and non-natural amino acids. Overall, bundlers provide a simple, versatile toolbox for a wide range of materials design and refinement, all while harnessing the design possibilities and function afforded by biologically inspired peptides.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1683-4>.

1. Lutolf, M. P. & Hubbell, J. A. Synthetic biomaterials as instructive extracellular microenvironments for morphogenesis in tissue engineering. *Nat. Biotechnol.* **23**, 47–55 (2005).
2. Egelman, E. H. et al. Structural plasticity of helical nanotubes based on coiled-coil assemblies. *Structure* **23**, 280–289 (2015).
3. Yang, Y. J., Holmberg, A. L. & Olsen, B. D. Artificially engineered protein polymers. *Annu. Rev. Chem. Biomol. Eng.* **8**, 549–575 (2017).
4. Seeman, N. C. & Sleiman, H. F. DNA nanotechnology. *Nat. Rev. Mater.* **3**, 17068 (2018).
5. Ding, Z. Z. et al. Simulation of ECM with silk and chitosan nanocomposite materials. *J. Mater. Chem. B* **5**, 4789–4796 (2017).
6. Dooling, L. J., Buck, M. E., Zhang, W.-B. & Tirrell, D. A. Programming molecular association and viscoelastic behavior in protein networks. *Adv. Mater.* **28**, 4651–4657 (2016).



7. Li, L. Q., Tong, Z. X., Jia, X. Q. & Kiick, K. L. Resilin-like polypeptide hydrogels engineered for versatile biological function. *Soft Matter* **9**, 665–673 (2013).
8. Bryson, J. W. et al. Protein design—a hierarchical approach. *Science* **270**, 935–941 (1995).
9. Huang, P.-S. et al. High thermodynamic stability of parametrically designed helical bundles. *Science* **346**, 481–485 (2014).
10. Zhang, H. V. et al. Computationally designed peptides for self-assembly of nanostructured lattices. *Sci. Adv.* **2**, e1600307 (2016).
11. Dogic, Z. & Fraden, S. Ordered phases of filamentous viruses. *Curr. Opin. Colloid Interface Sci.* **11**, 47–55 (2006).
12. Dogic, Z. Surface freezing and a two-step pathway of the isotropic-smectic phase transition in colloidal rods. *Phys. Rev. Lett.* **91**, 165701 (2003).
13. Gittes, F., Mickey, B., Nettleton, J. & Howard, J. Flexural rigidity of microtubules and actin-filaments measured from thermal fluctuations in shape. *J. Cell Biol.* **120**, 923–934 (1993).
14. Zhang, R., Kumar, N., Ross, J. L., Gardel, M. L. & de Pablo, J. J. Interplay of structure, elasticity, and dynamics in actin-based nematic materials. *Proc. Natl Acad. Sci. USA* **115**, E124–E133 (2018).
15. Zhao, J. G. et al. Freeze-thaw cycling induced isotropic-nematic coexistence of amyloid fibrils suspensions. *Langmuir* **32**, 2492–2499 (2016).
16. Bharti, B., Fameau, A. L., Rubinstein, M. & Velev, O. D. Nanocapillarity-mediated magnetic assembly of nanoparticles into ultraflexible filaments and reconfigurable networks. *Nat. Mater.* **14**, 1104–1109 (2015).
17. Hume, J. et al. Engineered coiled-coil protein microfibers. *Biomacromolecules* **15**, 3503–3510 (2014).
18. Papapostolou, D. et al. Engineering nanoscale order into a designed protein fiber. *Proc. Natl Acad. Sci. USA* **104**, 10853–10858 (2007).
19. Dong, H., Paramonov, S. E. & Hartgerink, J. D. Self-assembly of alpha-helical coiled coil nanofibers. *J. Am. Chem. Soc.* **130**, 13691–13695 (2008).
20. Usov, I., Adamcik, J. & Mezzenga, R. Polymorphism complexity and handedness inversion in serum albumin amyloid fibrils. *ACS Nano* **7**, 10465–10474 (2013).
21. Usov, I. & Mezzenga, R. FiberApp: an open-source software for tracking and analyzing polymers, filaments, biomacromolecules, and fibrous objects. *Macromolecules* **48**, 1269–1280 (2015).
22. VandenAkker, C. C., Engel, M. F. M., Velikov, K. P., Bonn, M. & Koenderink, G. H. Morphology and persistence length of amyloid fibrils are correlated to peptide molecular structure. *J. Am. Chem. Soc.* **133**, 18030–18033 (2011).
23. Aggeli, A. et al. Hierarchical self-assembly of chiral rod-like molecules as a model for peptide  $\beta$ -sheet tapes, ribbons, fibrils, and fibers. *Proc. Natl Acad. Sci. USA* **98**, 11857–11862 (2001).
24. Manning, G. S. The persistence length of DNA is reached from the persistence length of its null isomer through an internal electrostatic stretching force. *Biophys. J.* **91**, 3607–3616 (2006).
25. Averick, S. et al. Cooperative, reversible self-assembly of covalently pre-linked proteins into giant fibrous structures. *Angew. Chem. Int. Ed.* **53**, 8050–8055 (2014).
26. Oshaben, K. M. & Horne, W. S. Tuning assembly size in peptide-based supramolecular polymers by modulation of subunit association affinity. *Biomacromolecules* **15**, 1436–1442 (2014).
27. Thomas, F., Burgess, N. C., Thomson, A. R. & Woolfson, D. N. Controlling the assembly of coiled-coil peptide nanotubes. *Angew. Chem. Int. Ed.* **55**, 987–991 (2016).
28. Ok, J. M. et al. Control of periodic defect arrays of 8CB (4'-n-octyl-4-cyano-biphenyl) liquid crystals by multi-directional rubbing. *Soft Matter* **9**, 10135–10140 (2013).
29. Dogic, Z. Filamentous phages as a model system in soft matter physics. *Front. Microbiol.* **7**, 1013 (2016).
30. Huang, B., Bates, M. & Zhuang, X. Super-resolution fluorescence microscopy. *Annu. Rev. Biochem.* **78**, 993–1016 (2009).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

Schemes for synthesis and details of all amino acids, peptides and functionalized gold nanoparticles are described in detail in the Supplementary Information.

### Preparing bundlemer polymers by the thiol–maleimide reaction

**Rigid rods.** We prepared two separate solutions, one of peptide 1 and one of peptide 2 (Extended Data Fig. 1), each containing 1 mM peptide in phosphate buffer (pH 6, 25 mM), for the respective formation of each homomeric bundle. We then mixed equal volumes of the two bundle solutions and added 0.2 eq. of tris(2-carboxyethyl)phosphine (TCEP; 50 mM in distilled water). The bundle mixture was shaken overnight at room temperature to produce the rods.

**Alkyne-containing rods.** To make rods of peptide 6 and peptide 2 (Extended Data Fig. 1) for PEG conjugation, we added 1  $\mu$ mol of each peptide as freshly lyophilized powder to a 2-ml Eppendorf tube. To the solid mixture, we added 200  $\mu$ l of Millipore H<sub>2</sub>O to make a 5-mM solution of bundles, and vortexed the solution. The pH of the system was checked (the pH was roughly 6). The solution was mixed at room temperature for 12 h and at 60 °C for an additional 12 h.

**Representative maleimide-excessive short rigid rods.** We prepared separate solutions of peptide 1 and peptide 2 in phosphate buffer (pH 6, 25 mM) at 1% w/v. We mixed different volumes of the two bundle solutions to achieve a molar ratio of maleimide/thiol of 10/9 (for example, we could mix 0.528 ml of peptide-1 solution with 0.472 ml of peptide-2 solution to make a 1-ml 1% w/v solution). We then added TCEP (50 mM in distilled water) at 0.2 eq. relative to peptide 2. The mixture was shaken overnight at room temperature to produce the short rigid rods.

**Semi-flexible chains.** We prepared separate solutions of 1 mM peptide 1 and 0.25 mM PETMP (Extended Data Fig. 6; Sigma) in phosphate buffer (pH 6, 25 mM). We mixed the same volumes of the two solutions, and added 0.2 eq. of TCEP (50 mM in distilled water) relative to peptide 1. The mixture was shaken overnight at room temperature to produce the semi-flexible chains.

**Kinked chains.** We prepared 1 ml of the 1% w/v short rigid rod solution as described above (with an excess of maleimide-containing bundles in order to guarantee that the short rods had maleimide termini; the molar ratio of maleimide/thiol was 10/9). PETMP (2.9  $\mu$ l, 50 mM) was added to produce a maleimide/thiol ratio of 1. The mixture was shaken at room temperature for a week to produce kinked chains.

### Conjugation of azide-functionalized PEG to alkyne-functionalized rods

Stock solutions in Millipore H<sub>2</sub>O were prepared of CuSO<sub>4</sub> (200 mM), tris-hydroxypropyltriazolylmethylamine (THPTA; 400 mM) and sodium ascorbate (1 M). To a 1.5-ml Eppendorf tube, we added 40  $\mu$ l of 5-mM peptide, which had previously been assembled into rods. We then dissolved N<sub>3</sub>-PEG<sub>2000</sub> (2 mg; 5 eq.) in 160  $\mu$ l of Millipore water and added this to the peptide-rod solution. From the respective stock solutions we added 1  $\mu$ l of CuSO<sub>4</sub> (2 eq.), 1  $\mu$ l of THPTA (4 eq.) and 2  $\mu$ l of sodium ascorbate (10 eq.). The solution was purged with nitrogen and then vortexed and mixed at 40 °C for 48 h.

### Formation of gold-nanoparticle chains with peptide-7 bundles

We dissolved 1 mg of gold nanoparticles functionalized with maleimide groups (Supplementary Information scheme 7) in 1 ml dimethyl formamide (DMF) to make a transparent yellow solution. We added 5  $\mu$ l of DMF solution consisting of 1 mM thiol-functionalized peptide 7 (Extended Data Fig. 1) and 0.2 eq. TCEP to the gold-nanoparticle DMF solution. The mixture was reacted overnight at room temperature to produce gold

nanoparticles conjugated with peptide 7. The solution was then dialysed against pure water to afford coiled-coil bundle formation by the conjugated peptide 7, producing chains of nanoparticle–peptide-7 bundles.

### Representative hydrogel of peptide 1 with four-arm PEG tetrathiol

We dissolved 8 mg of peptide 1 and 10.86 mg of four-arm PEG tetrathiol (20 kDa; JenKem Technology) in 620  $\mu$ l phosphate buffer (pH 6, 25 mM). We added 0.2 eq. TCEP (9  $\mu$ l, 50 mM in distilled water) relative to peptide 1. The mixture was shaken at room temperature. The hydrogel usually formed after two to four hours of reaction.

### Transmission electron microscopy

Carbon-coated 200-mesh copper grids (CF200-Cu; Electron Microscopy Sciences) were freshly treated by glow discharge using a plasma cleaner (PDC-32G; Harrica Plasma) at low level for 20 s. We then dropped 4  $\mu$ l of sample solution on the grid. After 1 min, the remaining liquid was blotted from the edge of the grid using filter paper. The grid was allowed to air dry for 10 min before TEM on a FEI TALOS F200C microscope. For negative staining, we applied 6  $\mu$ l of an aqueous solution of phosphotungstic acid (2% w/v; pH 6) to the dried grid and incubated it for 15–30 s. The grid was then blotted with filter paper. The stained grid was allowed to sit for 10 min before TEM.

CryoTEM imaging was also performed on a FEI TALOS F200C microscope with accelerating voltage at 200 kV. We used lacey grids (Ted Pella) for all grid preparation after oxygen plasma treatment. We prepared vitrified grids for cryoTEM using the Vitrobot, an automated plunge-freezing device that vitrifies a thin solution layer to the temperature of liquid nitrogen. A sample droplet of 1.5  $\mu$ l was deposited onto the plasma-cleaned lacey grids. Depending on sample viscosity and concentration, we adjusted blotting parameters in order to obtain the optimal liquid-film thickness, usually requiring two to three blottings lasting 1–2 s at 100% humidity. After blotting, we allowed the sample grids to relax for 2 s. In order to achieve an extremely fast rate of cooling of the homogenous vitreous layer, we plunged the grid into liquid ethane (at roughly –175 °C) and then transferred it to liquid nitrogen for storage. During imaging, the cryoTEM holder was maintained at –177 °C to prevent ice crystallization or sublimation. The images were recorded with a FEI Ceta 16M (charge-coupled-device) or Falcon-II camera (CMOS) at a low electron dose.

### Small-angle neutron scattering

SANS experiments were performed on the NG-B 30-m SANS instrument, part of the Center for High Resolution Neutron Scattering (CHRNS) at the National Center for Neutron Research (NCNR), National Institute of Standards and Technology, Gaithersburg, MD, USA. Installed on the 60 mm  $\times$  60 mm split neutron guide NG-B, this instrument delivers a neutron beam of wavelength ( $\lambda$ ) approximately 6 Å, with a resolution ( $\Delta\lambda/\lambda$ ) of 10% at full width at half maximum (FWHM). The detector installed on this instrument is a 640 mm  $\times$  640 mm <sup>3</sup>He position-sensitive counter with a resolution of 5.08 mm  $\times$  5.08 mm. We used sample-to-detector lengths of 1 m, 4 m and 13 m to cover a  $q$ -range of 0.0035 Å<sup>–1</sup> to 0.4 Å<sup>–1</sup> for scattering experiments involving the 1% w/v semi-flexible chains. (Here,  $q$  is the scattering vector given by  $q = 4\pi \sin(\theta/2)/\lambda$ , where  $\theta$  is the scattering angle and  $\lambda$  is the neutron wavelength.) A wider  $q$ -range of 0.0015 Å<sup>–1</sup> to 0.35 Å<sup>–1</sup> was covered for scattering experiments on 1% w/v rigid rods, enabled by the additional use of neutron lenses at the 13-m detector configuration. We corrected the raw data obtained from scattering experiments for background noise and radiation, detector sensitivity, and open beam transmission using IgorPro<sup>31</sup> software to obtain a normalized scattering intensity curve. We calculated standard deviations statistically using the number of averaged detector counts at each data point. The reduced 1D scattering intensity obtained after buffer subtraction was fitted to various models using SasView<sup>32</sup> software.

Scattering from an isotropic solution of non-interacting, monodisperse species is described by the general equation<sup>33</sup>:

$$I(q) = nV^2(\Delta\rho)^2 P(q)$$

Here,  $I(q)$  is the normalized scattered intensity as a function of the scattering vector  $q$ ;  $n$  is the number density of scattering species;  $V$  is the volume of each scatterer;  $\Delta\rho$  is the difference in scattering length density (SLD) between the scattering species and solvent; and  $P(q)$  is the form factor, given by the average geometric shape of the scattering species in solution.

To fit the scattering curve from the rigid-rod solution in SasView, we chose a cylinder model, the  $P(q)$  of which is calculated by<sup>34,35</sup>:

$$P(q) = \frac{\text{scale}}{V} \int_0^{\pi/2} f^2(q, \alpha) \sin \alpha d\alpha + \text{background}$$

where

$$f(q, \alpha) = 2(\Delta\rho)V \times \frac{\sin(qL\cos\frac{\alpha}{2})}{(qL\cos\frac{\alpha}{2})} \times \frac{J_1(qr\sin\alpha)}{qr\sin\alpha}$$

Here,  $J_1$  is the first-order Bessel function;  $\alpha$  is the angle between the cylinder axis and the scattering vector  $q$ ;  $L$  is the length of the cylinder; and  $r$  is its radius; 'scale' is the volume fraction of cylinders; and 'background' is the background scattering intensity. An integral over  $\alpha$  from 0 to  $\pi/2$  radians averages the scattering intensity over all possible orientations of rods in an isotropic solution.

For semi-flexible rods, a flexible cylinder model fit was performed in SasView. Its form factor  $P(q)$  is defined by the equation<sup>31,33,34</sup>:

$$P(q) = \frac{\text{scale}}{V} f(q, \alpha)^2 + bkg$$

Where  $f(q, \alpha)^2$  is a squared average scattering over all possible orientations  $\alpha$  for a given scattering vector  $q$ . A worm-like semi-flexible cylindrical chain is used with a contour length  $L$ , radius  $r$  and Kuhn length,  $K_p$ , of  $2l_p$ , where  $l_p$  is the persistence length of the worm-like chain<sup>34</sup>. This model also incorporates excluded-volume interactions between segments of the worm-like chain in solution.

## Optical microscopy

We prepared a 15% w/v solution of short rigid rods with an average length of roughly 200 nm by gently concentrating a dilute solution of rods as in ref. <sup>36</sup>. Nitrogen gas was blown at a low flow rate into a flask containing a 0.5% w/v solution of short rigid rods under constant mild agitation. This slowly concentrated the solution by evaporating water molecules from the exposed air–water interface, avoiding the formation of inhomogeneities and nonequilibrium structures such as crystalline aggregates during the concentration process. The concentrated solution was then adjusted to pH 2 by adding a few drops of 1 M hydrochloric acid. We added anhydrous sodium chloride to yield a rod solution containing the salt at roughly 100 mM. Samples for polarized optical microscopy were prepared by adding 2  $\mu$ l of the solution between a clean glass slide and a cover slip. We immediately investigated the freshly prepared sample slides for birefringence resulting from the formation of liquid-crystalline phases under polarized light in transmission mode on an Olympus BX60 light microscope at 20 °C. High-resolution images were captured by a Nikon DS-Fi1 digital camera and the images were analysed using NIS-Elements imaging software.

## Rheological measurements

Rheological measurements were performed on a TA Instruments DHR-3 rheometer (TA Instruments). A hydrogel (160  $\mu$ l) was deposited onto the rheometer stage. A 20-mm stainless-steel parallel plate was used and

the gap height was set to 500  $\mu$ m for measurement. Oil was applied to seal the sample. We monitored the storage modulus  $G'$  and loss modulus  $G''$  under an applied strain of 0.01% to 10,000% at a frequency of 1 rad s<sup>-1</sup> for the strain sweep, and a frequency of 0.1 rad s<sup>-1</sup> to 200 rad s<sup>-1</sup> at strain 0.1% for the frequency sweep. Temperature sweeps were performed from 25 °C to 80 °C at 5° increments. Temperature-reversible experiments were carried out by subjecting the gel to a strain of 0.5% and frequency of 1 rad s<sup>-1</sup>.

## Stochastic optical reconstruction microscopy

**Preparation of solutions of rigid rods with red or green individual rod segments.** We prepared two 0.5-ml solutions of short rigid rods with thiol termini through click reactions between identical volumes of peptides 2 (1 mM) and 3 (0.9 mM), or between identical volumes of peptides 2 (1 mM) and 4 (0.9 mM) (Extended Data Fig. 1). We then mixed each of the two rigid-rod solutions with an identical volume containing the appropriate amount of peptide 1 (100  $\mu$ l, 1 mM), to produce a ratio of total added maleimide groups to thiol end groups of 1/1 in the entire solution. The mixture was shaken at room temperature for a week to produce longer chains containing short rod segments of red or green dye.

## Denaturation and reassembly of rigid rods with mixed red and green segments

We heated a solution of rigid rods with red- and green-dye-containing segments to 90 °C for 10 min in order to denature the rods. We then incubated the solution at 4 °C for 24 h in order to reassemble the constituent bundles, resulting in rigid rods with green and red dye mixed along the entire length of the rods.

**STORM imaging.** STORM images were taken on a Zeiss Elyra PS.1 super-resolution microscope. Rods were mounted on a high-precision 22 mm  $\times$  22 mm coverslip (Zeiss) by applying 10  $\mu$ l of a rod solution for 10 s. The remaining liquid was removed using filter paper. The sample adhered to the cover slip was rinsed with phosphate buffer (pH 6, 25 mM) five times. We added an oxygen-scavenging buffer (540 mM glucose, 3.1  $\mu$ M catalase, 7.6  $\mu$ M glucose oxidase, 10 mM NaCl, 20 mM cysteamine) in 58 mM Tris-HCl to the sample just before image acquisition, and sealed it in a magnetic conflat flange chamber (Chamlyde). STORM images with 4-chloro-7-nitrobenzofurazan or 5(6)-carboxy-tetramethylrhodamine were taken with a Plan-Apochromat 100 $\times$ /1.46 oil objective with 488-nm or 561-nm laser excitation, respectively. For each STORM image, we acquired 500 frames, aligned them using a model-based algorithm, and filtered them with a precision of 1–30 nm. For STORM imaging of rigid rods with individual rod segments containing either red or green dye, images were taken sequentially, merged, and then aligned. All image-processing steps were completed with Zen 2012 software.

## Atomic force microscopy

We carried out AFM with a Bruker multimode system, using Bruker ScanAsyst Air ultrasharp tips with a nominal tip radius of 2 nm and a spring constant of 0.4 N m<sup>-1</sup>. We used freshly cleaved mica discs (Ted Pella) as substrates for sample deposition. Samples were prepared by spin-coating 20  $\mu$ l of a roughly 5 mM sample of PEG<sub>2000</sub>-conjugated bundlemer rods onto a freshly cleaved mica disc using a spin-coater (WS-650SZ, Laurel Technologies). In a first stage the spin-coater was set to 500 r.p.m. for 10 s, during which the sample solution was pipetted onto the substrate. The speed was then increased sequentially to 2,000 r.p.m. and held for 3 min at the second stage to spin off the liquid. During the 3 min of spin coating, 20  $\mu$ l of Milli-Q water was applied to wash off excess salts. The mica disc appeared to be dry after the spin-coating and was rested for at least 1 h before being subjected to AFM imaging. The instrument was operated in peak force tapping mode. A new AFM tip was used for each different sample. The peak force set point was adjusted manually in order to optimize the spatial resolution as well as to minimize sample damage. Scanning was performed in the horizontal direction and repeated in the 45° diagonal direction to exclude



scanning artefacts. Micrographs were recorded digitally using Bruker Nanoscope software with 512 lines at a 0.5-Hz scan rate and corrected for background undulations using the in-software algorithm function.

## Data availability

The data supporting the findings of this study are available within the paper and its Supplementary Information files.

31. Kline, S. R. Reduction and analysis of SANS and USANS data using IGOR Pro. *J. Appl. Cryst.* **39**, 895–900 (2006).
32. SasView. <http://www.sasview.org>
33. Guinier, A. & Fournet, G. *Small-Angle Scattering of X-Rays* (Wiley, 1955).
34. Pedersen, J. S. & Schurtenberger, P. Scattering functions of semiflexible polymers with and without excluded volume effects. *Macromolecules* **29**, 7602–7612 (1996).
35. Chen, W., Butler, P. D. & Magid, L. J. Incorporating intermicellar interactions in the fitting of SANS data from cationic wormlike micelles. *Langmuir* **22**, 6539–6548 (2006).
36. Jung, J. & Mezzenga, R. Liquid crystalline phase behavior of protein fibers in water: experiments versus theory. *Langmuir* **26**, 504–514 (2010).
37. Haider, M. J., Zhang, H., Sinha, N., Saven, J. G. & Pochan, D. J. Self-assembly and soluble aggregate behavior of computationally designed coiled-coil peptide bundles. *Soft Matter* **14**, 5488–5496 (2018).
38. Käs, J., Strey, H. & Sackmann, E. Direct imaging of reptation for semiflexible actin filaments. *Nature* **368**, 226–229 (1994).
39. Boal, D. H. *Mechanics of the Cell* (Cambridge Univ. Press, 2012).
40. Falvo, M. R. et al. Manipulation of individual viruses: friction and mechanical properties. *Biophys. J.* **72**, 1396–1403 (1997).

**Acknowledgements** Primary funding was provided by the Department of Energy, Office of Basic Energy Sciences, Biomolecular Materials Program under grants DE-SC0019355 and DE-SC0019282. D.J.P. and N.S. acknowledge stipend support and support for neutron-scattering experiments under cooperative agreements 70NANB12H239 and 70NANB17H302 from the National Institute of Standards and Technology (NIST), US Department of Commerce. D.J.P.

and N.S. also acknowledge the support of the NIST in providing neutron research facilities. Neutron facilities are supported in part by the National Science Foundation (NSF) under agreement DMR-0944772. We acknowledge the support of the National Institutes of Health (NIH), grant RO1 EB006006, and the University of Delaware NIH Centers of Biomedical Research Excellence (COBRE) grants 1P30.GM110758 and 1P20.RR017716 for instrument resources. We acknowledge the Delaware IDeA Network of Biomedical Research Excellence grant P20 GM103446 for support of the Delaware Biotechnology Institute. J.G.S. acknowledges partial support from NSF CHE 1709518 and the Penn Laboratory for Research on the Structure of Matter (MRSEC; grant NSF DMR-1120901). D.J.P. and J.L. acknowledge stipend support from the NSF under grant DMREF-1629156. D.J.P., J.G.S., Y.T. and H.Z. acknowledge stipend support from the NSF under grants DMR-1234161 and DMR-1235084. The statements, findings, conclusions and recommendations herein are those of the authors and do not necessarily reflect the view of NIST, the US Department of Commerce, the US Department of Energy, or the University of Delaware. D.W., D.J.P. and C.J.K. acknowledge internal research support from the University of Delaware.

**Author contributions** D.J.P., C.J.K. and J.G.S. conceived of the project and mentored research activity, as well as contributed to the writing of the manuscript. D.W. carried out peptide synthesis, gold-nanoparticle synthesis and peptide conjugation with subsequent assembly into nanoparticle-bundle chains, and bundle and bundlemer polymer formation and characterization. D.W. also contributed to the writing of the manuscript. J.L. and N.S. contributed to characterization via TEM, cryoTEM, POM and SANS, with N.S. also contributing to the writing of the manuscript. B.P.S. and N.I.H. performed peptide synthesis, molecular characterization, rod formation and PEG conjugation. Y.T. performed AFM characterization. J.C. performed STORM characterization and contributed to the writing of the manuscript. H.V.Z. performed the computational design and modelling of peptides.

**Competing interests** The authors declare no competing interests.

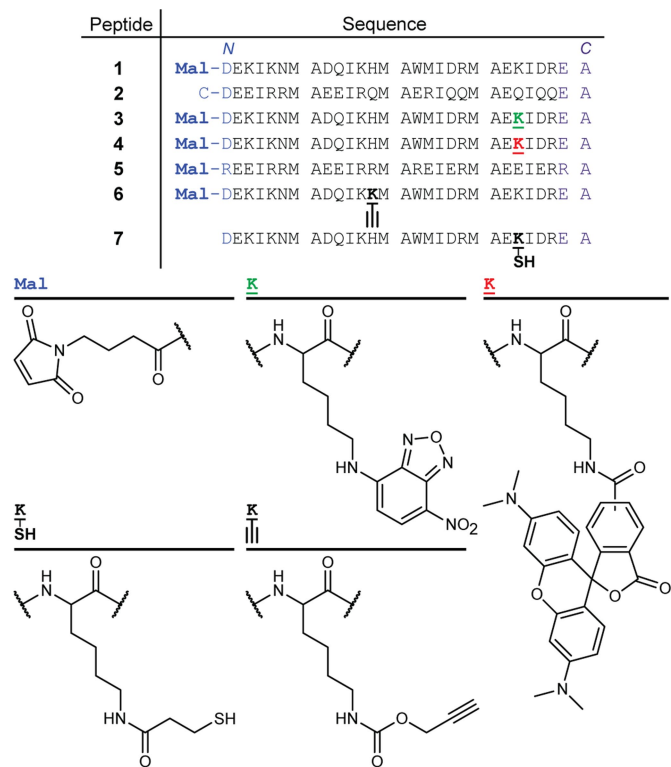
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1683-4>.

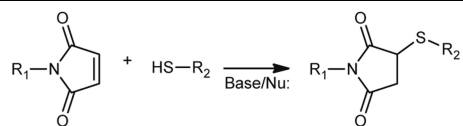
**Correspondence and requests for materials** should be addressed to J.G.S., C.J.K. or D.J.P.

**Peer review information** *Nature* thanks Vincent Conticello, Shuguang Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

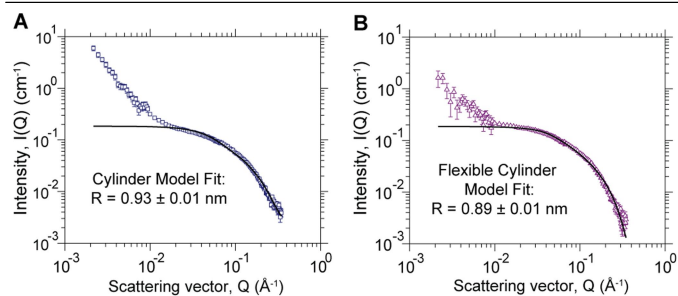


**Extended Data Fig. 1 | Structures of peptides 1–7.** The single-letter amino acid sequences of peptides 1–7 are shown at the top, noting the peptides in which maleimide (Mal) has been chemically added to the N terminus, and those in which lysine amino acids (K) have been modified with the structures shown at the bottom. The unmodified amino acid sequences of peptide 1, peptide 2 (without an N-terminal cysteine) and peptide 6 have previously been denoted P622\_6 (ref. <sup>10</sup>), BNDL1 (ref. <sup>10</sup>) and BNDL2 (ref. <sup>37</sup>), respectively.

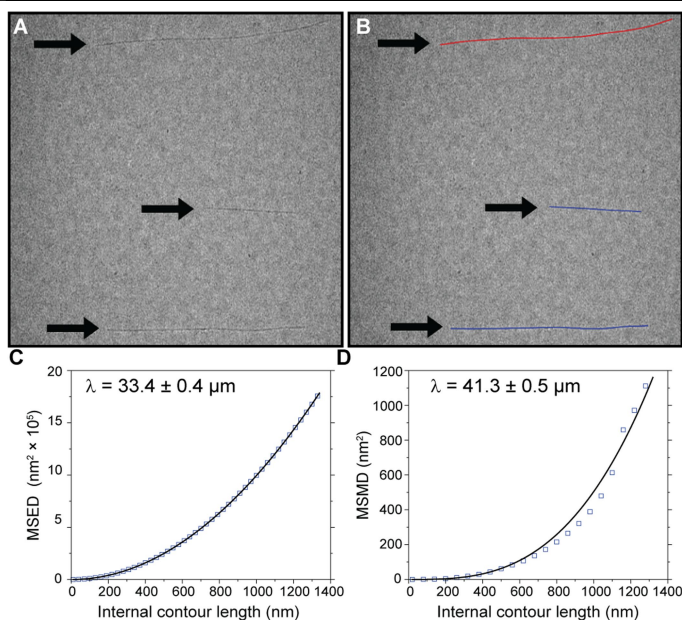


**Extended Data Fig. 2 | The thiol-maleimide click reaction.**  $\text{R}_1$  and  $\text{R}_2$  represent the possible remainders of the molecular structures. The thiol-maleimide reaction is catalysed by either a base or a nucleophile (Nu) catalyst.

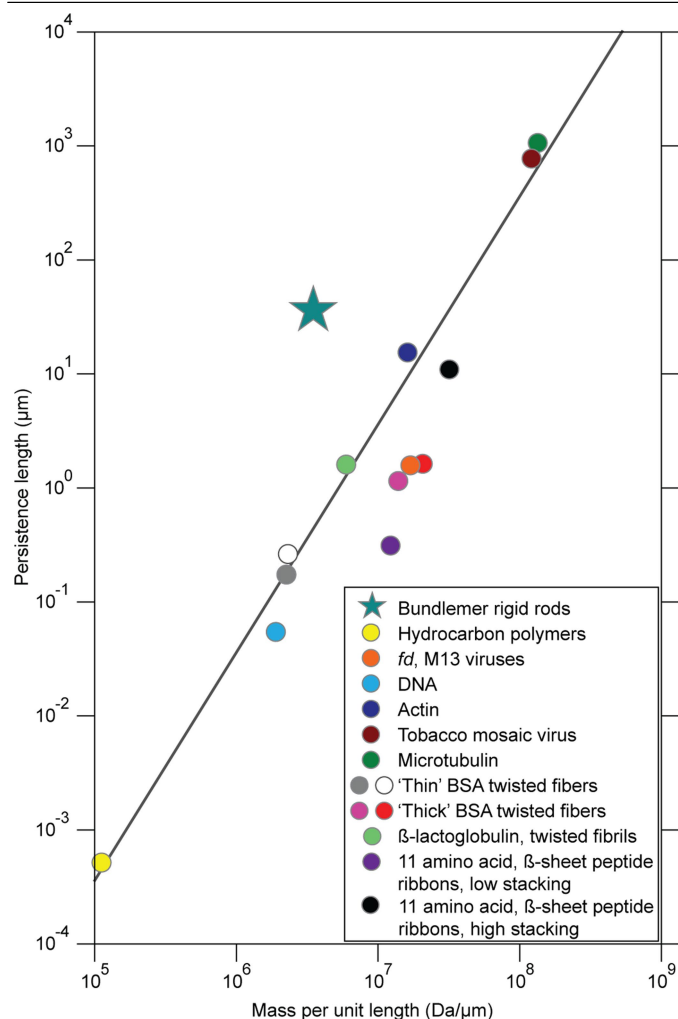




**Extended Data Fig. 3 | SANS of peptide rods with different linker chemistries. a,** Scattering from rigid rods (blue squares) and the corresponding rigid cylinder fit (black curve). Rigid rods (identical to those in Fig. 1b) were assembled in 25 mM pH 6 phosphate buffer prepared in deuterated water at 20 °C. **b,** Scattering from semi-flexible fibres (purple triangles) and the corresponding flexible cylinder fit (black curve). Semi-flexible chains (identical to those shown in Fig. 1e) were dissolved in 25 mM pH 6 phosphate buffer prepared in deuterated water at 20 °C. In each case,  $R$  is the fitted cylinder radius of the corresponding model.

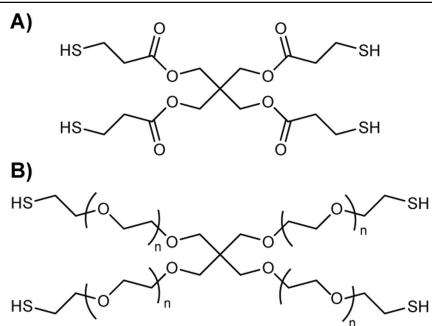


**Extended Data Fig. 4 | Estimation of the persistence length of rigid-rod bundle chains.** The estimation was carried out using FibreApp tracking and the analysis software of ref. <sup>21</sup>. **a**, CryoTEM of rigid-rod bundle chains. **b**, The same rigid-rod bundle chains after software tracking. The red trace is the last rod tracked with the software in the image and is used, along with the earlier blue traces, in the stiffness analyses. **c**, Plot of calculated mean-squared end-to-end distance (MSED) between contour segments along the tracked rod (blue squares), and the corresponding MSED fit (black curve) between contour segments. This fit is based on a worm-like chain model in two dimensions with the following theoretical dependence:  $\langle R^2 \rangle = 4\lambda[l - 2\lambda(1 - e^{-l/2\lambda})]$ , where  $\lambda$  is the persistence length and  $R$  is the direct distance between any pair of segments along a contour separated by arc length  $l$ . The persistence length is estimated to be  $33.4 \pm 0.4 \mu\text{m}$ . **d**, An alternative method for estimating the persistence length of very stiff, one-dimensional objects is the mean-squared midpoint displacement (MSMD). This plot shows the calculated MSMD between contour segments along the tracked rod (blue squares) and the corresponding MSMD fit (black curve). This fit is based on an equation that describes the behaviour of a midpoint deviation along a rod:  $\langle u_x^2 \rangle = l^3/48\lambda$ , where  $\langle u_x^2 \rangle$  is the MSMD between any pair of segments along a tracked-rod contour, separated by an arc length  $l$ , assuming that the displacements are small in comparison with the corresponding arc lengths ( $\langle u_x \rangle$  is much less than  $l$ ). This method estimates the persistence length of the rod to be  $41.3 \pm 0.5 \mu\text{m}$ .



**Extended Data Fig. 5 | Plot showing persistence length versus mass per unit length for various 1D polymers and molecular assemblies.** The plot is adapted with permission from ref. <sup>16</sup>, Springer Nature Limited. For peptide-bundlemer rigid rods, the persistence length was estimated from our cryoTEM data using the methods of ref. <sup>21</sup>. Other persistence lengths were taken from the following references: hydrocarbon polymers, ref. <sup>16</sup>; *fd* and M13 viruses, refs. <sup>11,12</sup>; DNA, refs. <sup>16,24</sup>; actin, refs. <sup>13,38</sup>; tobacco mosaic virus, refs. <sup>16,39,40</sup>; microtubulin, refs. <sup>13,39,40</sup>; thin (diameter 2–3 nm) and thick (diameter 4–6 nm) twisted bovine serum albumin (BSA) fibres, ref. <sup>20</sup>;  $\beta$ -lactoglobulin  $\beta$ -sheet-rich twisted fibrils (with diameters ranging from 1 nm to 6 nm, with a mean of roughly 3 nm), ref. <sup>11,22</sup>; amino acid  $\beta$ -sheet ribbons with low stacking (producing ribbon diameters of roughly 4 nm) and high stacking (producing ribbons with diameters of 4–8 nm), ref. <sup>23</sup>.

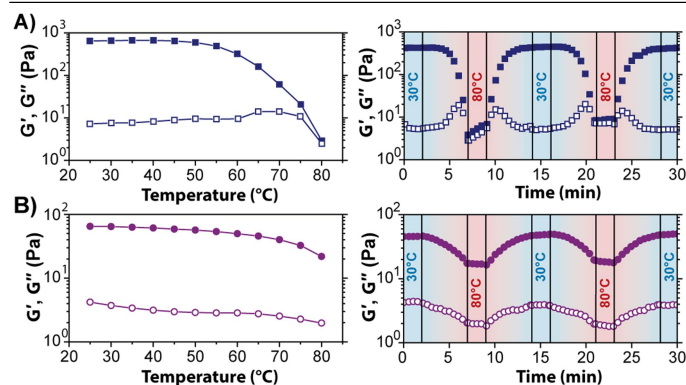




**Extended Data Fig. 6 | Chemical structures of bundlemer organic linkers.**

**a.** Structure of PETMP for the formation of semiflexible or kinked chains.

**b.** Chemical structure of four-arm PEG tetrathiol (20 kDa) for hydrogel formation.



**Extended Data Fig. 7 | Hydrogel network rheology.** Hydrogels were composed of peptides 1 or 5 (Extended Data Fig. 1) linked with four-arm PEG tetrathiol (Extended Data Fig. 6). We monitored the storage modulus  $G'$  (filled symbols) and loss modulus  $G''$  (open symbols) as functions of temperature (left) and as a result of the cycling of temperature (right); in the latter case, temperature ramps were performed over 5 min between isothermal measurements of 2-min duration. **a**, Peptide 1, with  $T_m = 55^{\circ}\text{C}$ . Peptide 1 produces a temperature-reversible hydrogel owing to the low melting temperature of the designed peptide. **b**, Peptide 5, with  $T_m = 85^{\circ}\text{C}$ . The hydrogel produced with peptide 5 is stable to approximately  $85^{\circ}\text{C}$  and shows much more rigid gel properties at all temperatures tested.

# Two-million-year-old snapshots of atmospheric gases from Antarctic ice

<https://doi.org/10.1038/s41586-019-1692-3>

Received: 6 July 2018

Accepted: 2 August 2019

Published online: 30 October 2019

Yuzhen Yan<sup>1\*</sup>, Michael L. Bender<sup>1,2</sup>, Edward J. Brook<sup>3</sup>, Heather M. Clifford<sup>4</sup>, Preston C. Kemeny<sup>1,5</sup>, Andrei V. Kurbatov<sup>4</sup>, Sean Mackay<sup>6</sup>, Paul A. Mayewski<sup>4</sup>, Jessica Ng<sup>7</sup>, Jeffrey P. Severinghaus<sup>7</sup> & John A. Higgins<sup>1</sup>

Over the past eight hundred thousand years, glacial–interglacial cycles oscillated with a period of one hundred thousand years (‘100k world’<sup>1</sup>). Ice core and ocean sediment data have shown that atmospheric carbon dioxide, Antarctic temperature, deep ocean temperature, and global ice volume correlated strongly with each other in the 100k world<sup>2–6</sup>. Between about 2.8 and 1.2 million years ago, glacial cycles were smaller in magnitude and shorter in duration (‘40k world’<sup>7</sup>). Proxy data from deep-sea sediments suggest that the variability of atmospheric carbon dioxide in the 40k world was also lower than in the 100k world<sup>8–10</sup>, but we do not have direct observations of atmospheric greenhouse gases from this period. Here we report the recovery of stratigraphically discontinuous ice more than two million years old from the Allan Hills Blue Ice Area, East Antarctica. Concentrations of carbon dioxide and methane in ice core samples older than two million years have been altered by respiration, but some younger samples are pristine. The recovered ice cores extend direct observations of atmospheric carbon dioxide, methane, and Antarctic temperature (based on the deuterium/hydrogen isotope ratio  $\delta D_{ice}$ , a proxy for regional temperature) into the 40k world. All climate properties before eight hundred thousand years ago fall within the envelope of observations from continuous deep Antarctic ice cores that characterize the 100k world. However, the lowest measured carbon dioxide and methane concentrations and Antarctic temperature in the 40k world are well above glacial values from the past eight hundred thousand years. Our results confirm that the amplitudes of glacial–interglacial variations in atmospheric greenhouse gases and Antarctic climate were reduced in the 40k world, and that the transition from the 40k to the 100k world was accompanied by a decline in minimum carbon dioxide concentrations during glacial maxima.

Earth has been cooling, and ice sheets expanding, over approximately the past 52 million years (Myr)<sup>11</sup>. Superimposed on this cooling are periodic changes in the Earth’s climate system driven by variations in the eccentricity (with periods of 400 and 100 kyr) and precession (23 and 19 kyr) of the Earth’s orbit around the Sun, and the tilt of the spin axis (about 41 kyr). From around 2.8 to 1.2 Myr ago (Ma), the Earth’s climate system oscillated between glacial and interglacial states with a period of about 40 kyr (the 40k world<sup>7</sup>). Between 1.2 and 0.8 Ma, an interval known as the ‘mid-Pleistocene transition’ (MPT), the period of glacial cycles lengthened to about 100 kyr and glacial periods became colder<sup>12</sup>, for reasons that are poorly understood. The post-MPT glacial cycles are characterized by a quasi-100-kyr period (the 100k world<sup>1</sup>).

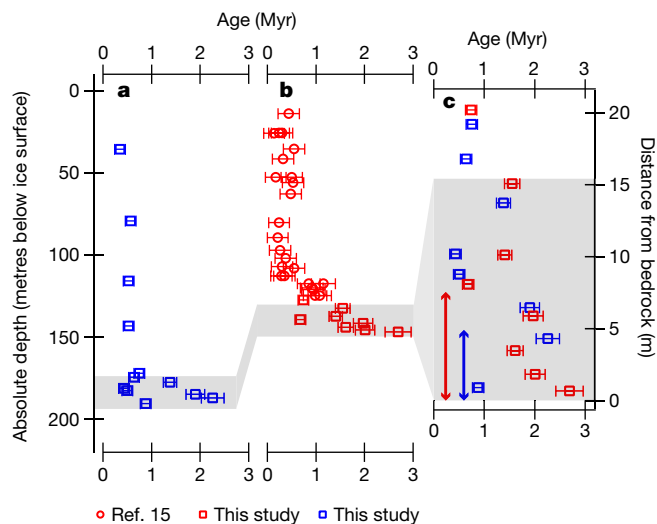
Studies of stratigraphically continuous ice cores have shown that atmospheric CO<sub>2</sub> is directly linked to Antarctic and global temperature over the last 800 kyr<sup>5,6</sup>. The coupling of the Earth’s climate and carbon cycle in earlier times, however, has not been conclusively demonstrated. Recent boron-based reconstructions have suggested that glacial CO<sub>2</sub>

declined across the MPT, but interglacial CO<sub>2</sub> remained stable<sup>8–10</sup>. These studies suggest that the range of CO<sub>2</sub> was reduced in the 40k world, as was the amplitude of glacial cycles. On the other hand, CO<sub>2</sub> estimates based on foraminiferal  $\delta^{13}C$  predict that CO<sub>2</sub> varied between 300 and 180 parts per million (ppm) in the 40k world<sup>13,14</sup>, hinting at a fundamentally different relationship between CO<sub>2</sub> and temperature. As both of these techniques rely on indirect measurements of atmospheric CO<sub>2</sub>, they have lower accuracy and precision (typically more than  $\pm 20$  ppm<sup>10</sup>; 2 s.d.( $\sigma$ )) than can be achieved by direct measurements in ice core samples ( $\pm 2$  ppm; 2 $\sigma$ ). Accurate and precise CO<sub>2</sub> measurements from ice cores that pre-date the MPT are therefore needed.

## Old ice records from the Allan Hills

Million-year-old ice was recently discovered<sup>15</sup> in shallow ice cores drilled in the Allan Hills Blue Ice Area (BIA), Antarctica (–76.73° N, 159.36° E; Extended Data Figs. 1, 2). In the Allan Hills, bedrock topography and

<sup>1</sup>Department of Geosciences, Princeton University, Princeton, NJ, USA. <sup>2</sup>School of Oceanography, Shanghai Jiao Tong University, Shanghai, China. <sup>3</sup>College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA. <sup>4</sup>Climate Change Institute, University of Maine, Orono, ME, USA. <sup>5</sup>Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA. <sup>6</sup>Department of Earth and Environment, Boston University, Boston, MA, USA. <sup>7</sup>Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA. \*e-mail: yuzheny@princeton.edu



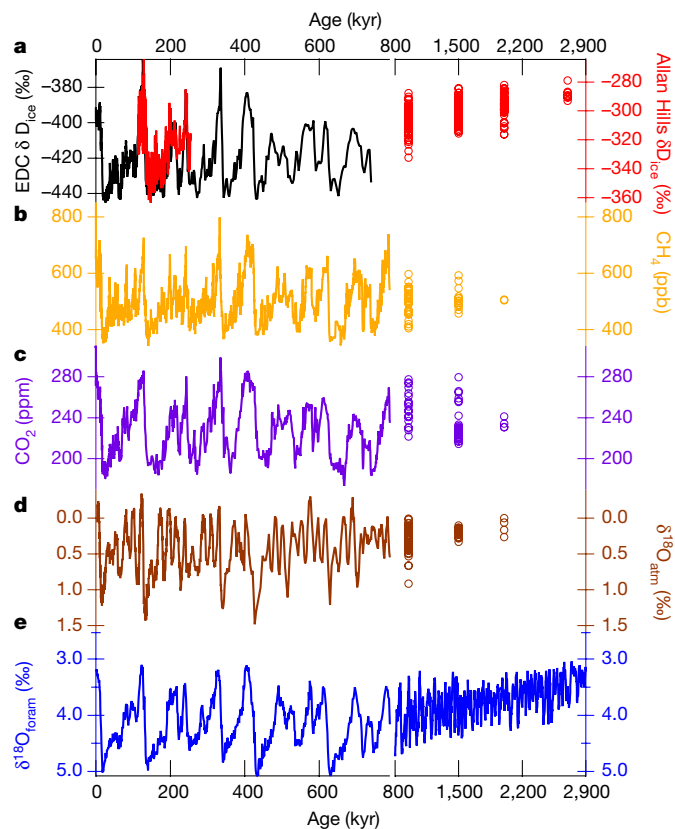
**Fig. 1 | Age–depth profile of Allan Hills ice cores. a**, ALHIC1502 data in blue; **b**, ALHIC1503 data in red. **c**,  $^{40}\text{Ar}_{\text{atm}}$  ages plotted against distance above bedrock with the same colour coding as in **a**, **b**. Data in circles were measured at Princeton University<sup>15</sup>. New measurements made at Scripps Institution of Oceanography are plotted as squares. Error bars represent the external reproducibility ( $1\sigma$ ; 110 kyr and 220 kyr for samples measured at Scripps and Princeton, respectively) of the measurement or 10% of the sample age, whichever is greater. Shading highlights sections in which ice that is more than 1 Myr old is present. Sections affected by respiration are marked by the vertical arrows (blue, ALHIC1502; red, ALHIC1503).

strong katabatic winds lead to the exhumation of old ice from depth to the surface<sup>16</sup>. In 2015–2016, two additional cores were drilled. The first extended the drilling to a depth of 147 m at the BIT-58 borehole (named ALHIC1503 hereafter), where approximately 1-Myr-old ice had previously been discovered. The second (ALHIC1502), drilled 56 m to the southwest, reached bedrock at 191 m. These ice cores are dated by measuring the  $^{40}\text{Ar}$  deficit in trapped air relative to the modern atmosphere<sup>17</sup> ( $^{40}\text{Ar}_{\text{atm}}$ ), and represent an archive of the Earth's climate extending well into the 40k world. A nearby blue-ice core (Allan Hills Site 27; S27) gives a continuous record from about 100 to 250 ka, and serves as a reference section for comparison with ALHIC1502 and ALHIC1503<sup>18</sup>.

Age–depth profiles of the ALHIC1502 and ALHIC1503 cores (Fig. 1) show that both sites contain 300–500-kyr-old ice, overlying a basal unit confined to approximately the bottom 30 m, with ages that range from more than 1 to 2.7 Myr. At ALHIC1503, the deepest sample lies within about 1 m of bedrock and has the oldest  $^{40}\text{Ar}_{\text{atm}}$  age,  $2.7 \pm 0.3$  ( $1\sigma$ ) Myr. However, stratigraphic disturbance in both ALHIC1502 and ALHIC1503 is evident from abrupt age discontinuities (Fig. 1), accompanied by large swings in  $\delta^{18}\text{O}$  of  $\text{O}_2$  ( $\delta^{18}\text{O}_{\text{atm}}$ ) and  $\delta\text{D}_{\text{ice}}$  values<sup>15</sup>. In light of these complications, we interpret the paleoclimate records from the Allan Hills BIA as discrete ‘snapshots’—discontinuous but accurate portraits—of the climate state, rather than continuous time series.

### Climate properties in the old ice

Paleoclimate reconstructions from stratigraphically discontinuous ice sections face several challenges. First, there is the potential for sampling bias in the preserved climate states due to possible differences in accumulation rate and/or ice rheology. Second, diffusion in million-year-old ice may lead to smoothing of paleoclimate properties or proxies<sup>19,20</sup>, though this effect is likely to be minor in Allan Hills ice (see Methods). Third, discrete sampling might not capture the complete glacial–interglacial range of properties of interest. Overall, we estimate that  $76 \pm 14\%$  ( $2\sigma$ ) of the total atmospheric  $\text{CO}_2$  variability in the 40k world is recovered by our  $\text{CO}_2$  samples. This estimate is based upon our deductions that 90% of the true glacial–interglacial climate variability is



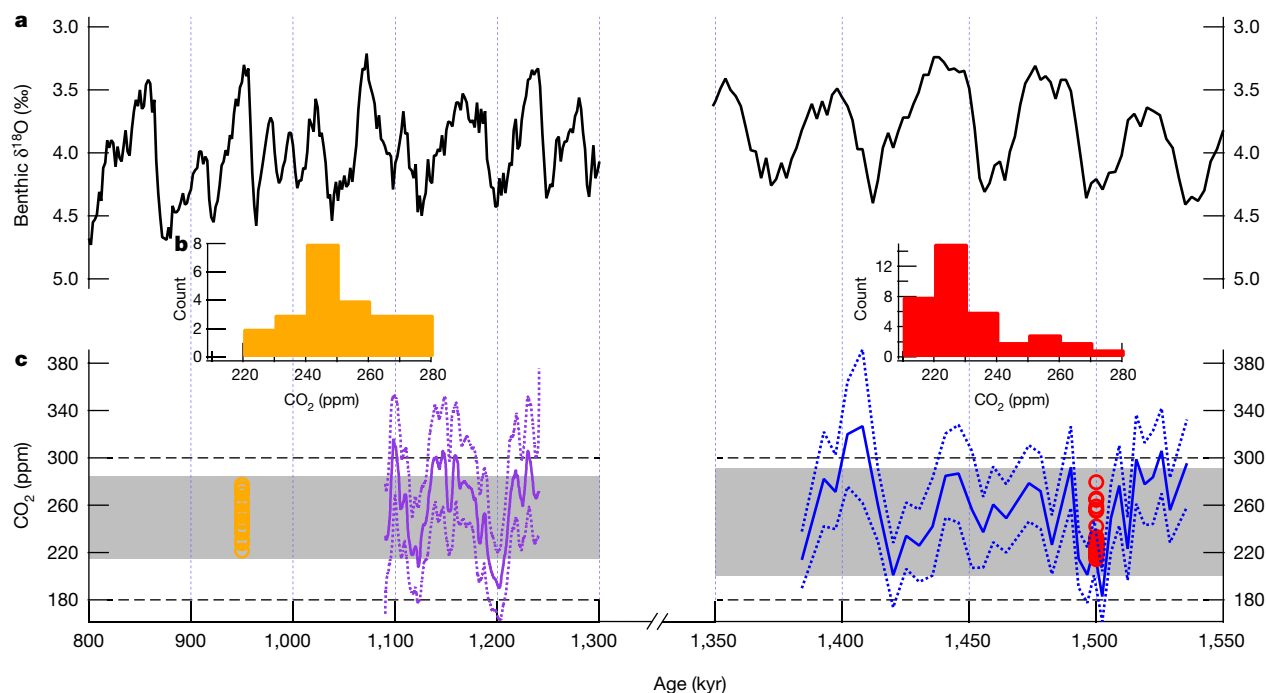
**Fig. 2 | Climate properties over the past 2.9 Myr documented in ice core and benthic foram records. a**, The continuous 800-kyr  $\delta\text{D}_{\text{ice}}$  record from the European Project for Ice Coring in Antarctica, Dome C (EDC) ice core<sup>4</sup> (black line); the continuous S27  $\delta\text{D}_{\text{ice}}$  record covering 120–250 ka<sup>18</sup> (red line); and the discrete Allan Hills  $\delta\text{D}_{\text{ice}}$  ice samples (red circles). **b**, The 800-kyr ice core  $\text{CH}_4$  record<sup>26</sup> (orange line) and the binned Allan Hills  $\text{CH}_4$  data (orange circles). **c**, The 800-kyr ice core  $\text{CO}_2$  record<sup>3,5,6,28</sup> (purple line) and the binned Allan Hills  $\text{CO}_2$  data (purple circles). Note that there are no reliable  $\text{CO}_2$  and  $\text{CH}_4$  analyses in the 2.7-Ma bin (see Methods). **d**, The 800-kyr ice core  $\delta^{18}\text{O}_{\text{atm}}$  record<sup>29,30</sup> (brown line) and binned values of Allan Hills  $\delta^{18}\text{O}_{\text{atm}}$  samples (brown circles). All  $\delta^{18}\text{O}_{\text{atm}}$  values are normalized to the modern atmosphere. **e**, The globally distributed benthic oxygen isotope stack over the Plio-Pleistocene (LR04)<sup>2</sup> showing decreased glacial–interglacial variability and less extreme glacials before 800 ka.

preserved in the Allan Hills ice, and that  $84 \pm 15\%$  ( $2\sigma$ ) of  $\text{CO}_2$  variability within the ice is captured by discrete samples, assuming a linear relationship between  $\text{CO}_2$  and benthic  $\delta^{18}\text{O}$  (see Methods).

Measured climate proxies and ice core properties across time (Fig. 2) include proxy records of regional temperature ( $\delta\text{D}_{\text{ice}}$ ),  $\text{CH}_4$ ,  $\text{CO}_2$ , and global  $\text{O}_2$  fractionation ( $\delta^{18}\text{O}_{\text{atm}}$ )<sup>21</sup>. Samples older than 800 kyr are assigned the age of the nearest  $^{40}\text{Ar}_{\text{atm}}$ -dated sample and then binned into two age groups: the MPT (800–1,200 kyr), and the 40k world (1.2–2.7 Myr). Within the 40k world bin, samples are further subdivided into three groups with average ages of 1.5, 2.0, and 2.7 Myr (see Methods and Fig. 2).

Improbably high  $\text{CO}_2$  concentrations or  $\delta^{13}\text{C}$  values of  $\text{CO}_2$  indicating contributions from respiratory  $\text{CO}_2$  were present in all 2.7-Myr-old samples, common among 2.0-Myr-old samples, but absent from 1.5-Myr-old samples (see Methods). We chose  $-7\text{‰}$  as the cut-off value for  $\delta^{13}\text{C}$  and rejected all  $\text{CO}_2$  samples below the shallowest  $\delta^{13}\text{C}$  sample with an isotopic value of lighter than  $-7\text{‰}$ . The application of these criteria excluded 21  $\text{CO}_2$  samples from our analysis.  $\text{CH}_4$  and  $\delta^{18}\text{O}_{\text{atm}}$  samples were also rejected when  $\delta^{13}\text{C}$  or  $\text{CO}_2$  measurements indicated that these properties were compromised. We continued to plot  $\delta\text{D}_{\text{ice}}$  values of the ice affected by respiration in Fig. 2 in the 2.0 and 2.7 Myr age groups, but excluded  $\delta\text{D}_{\text{ice}}$  data from anomalous  $\text{CO}_2$  samples in the 40k world.





**Fig. 3 | Comparison of blue-ice CO<sub>2</sub> record and boron-based CO<sub>2</sub> reconstructions during and before the MPT. a, LR04 benthic oxygen isotope stack<sup>2</sup>. b, Distribution of MPT CO<sub>2</sub> (left, orange) and pre-MPT CO<sub>2</sub> (right, red) data observed in Allan Hills ice. c, Comparison between ice core CO<sub>2</sub> (orange and red circles) and δ<sup>11</sup>B-based CO<sub>2</sub> reconstructions (purple and blue solid lines);**

dashed lines represent the 95% confidence intervals<sup>9,10</sup>. Shading represents the expanded range estimate, given the possibility that discrete Allan Hills CO<sub>2</sub> samples may not fully capture the true range of CO<sub>2</sub> variability. Black dashed lines represent the glacial–interglacial range of CO<sub>2</sub> in the 100k world<sup>3</sup>.

Ice dating back to about 100–250 ka, sampled at a nearby site<sup>18</sup>, records a δD<sub>ice</sub> range of –280 to –360‰ in the 100k world. δD<sub>ice</sub> values ranged from –284 to –316‰ in 40k-world ice and from –288 to –332‰ in MPT ice. Maximum values of δD<sub>ice</sub> were similar during all three time periods, lying within an 8‰ range. However, minimum values were higher in the MPT and the 40k world by at least 28‰. This implies that interglacial periods during the 40k world in Antarctica were not significantly warmer than those in the 100k world, but that glacial maxima were lower after the MPT.

CO<sub>2</sub> concentrations ranged from 221 to 277 ppm ( $n = 23$ )<sup>15</sup> in MPT ice, and from 214 to 279 ppm ( $n = 37$ ) in ice from the 40k world. Given that  $76 \pm 14\%$  ( $2\sigma$ ) of total CO<sub>2</sub> variability was captured by 37 discrete samples, and assuming that glacial and interglacial extremes are equally absent, the expected true range of 40k world CO<sub>2</sub> is  $86^{+19}_{-13}$  ppm ( $2\sigma$ ; Fig. 3). Our best estimate of 40k-world CO<sub>2</sub> concentrations is therefore 204–289 ppm. These ranges indicate that interglacial CO<sub>2</sub> concentrations in the 40k and 100k worlds were similar, whereas glacial CO<sub>2</sub> concentrations are likely to have been 24 ppm higher in the 40k world than in the 100k world. By comparison, reconstructions of atmospheric CO<sub>2</sub> from individual boron isotope measurements in foraminifera ranged from 190 to 320 ppm (Fig. 3), with the average extreme values being 234 and 277 ppm<sup>9,10</sup>.

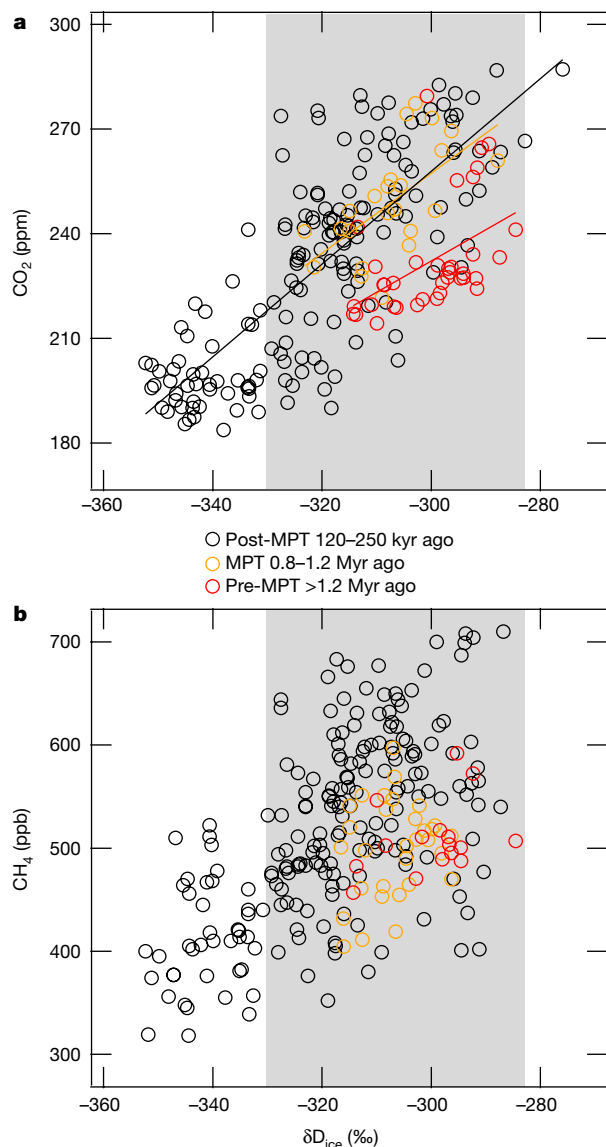
## Implications for climate evolution

Our findings appear to be inconsistent with hypotheses that attribute the transition into the 100k world to a long-term decline in both interglacial and glacial atmospheric CO<sub>2</sub><sup>22,23</sup>. Our data instead support hypotheses that link the MPT to greater ice sheet size and CO<sub>2</sub> drawdown during glacial maxima. Those hypotheses include, but are not limited to, enhanced dust delivery to the southern ocean<sup>24</sup> and changes in global ocean circulation<sup>25</sup> that resulted in an additional drawdown of atmospheric CO<sub>2</sub>, thereby enhancing the growth of continental ice sheets and leading to the emergence of 100-kyr glacial cycles.

Compared to CO<sub>2</sub> concentrations and δD<sub>ice</sub>, CH<sub>4</sub> concentrations and δ<sup>18</sup>O<sub>atm</sub> values in ice from the MPT and 40k world exhibited even less

glacial–interglacial variability. CH<sub>4</sub> concentrations ranged from 405 to 569 parts per billion (ppb) in MPT ice ( $n = 31$ ) and 457 to 592 ppb in 40k-world ice ( $n = 16$ ). These ranges are only about 40% of the range in the 100k world (approximately 350 to 800 ppb)<sup>26</sup>. Undersampling may play a role in this reduction, as CH<sub>4</sub> maxima in the 100k world have short durations<sup>26</sup>. In addition, discrete CH<sub>4</sub> measurements require more ice than CO<sub>2</sub> measurements (60 g versus 10 g). They therefore average over a longer interval of time than CO<sub>2</sub> samples. δ<sup>18</sup>O<sub>atm</sub> values ranged from 0.01‰ to 0.92‰ in MPT ice ( $n = 63$ ) and from 0.00‰ to 0.33‰ in 40k-world ice ( $n = 29$ ). In comparison, the range is from –0.30‰ to +1.48‰ in ice from the 100k world<sup>21</sup>. The δ<sup>18</sup>O<sub>atm</sub> values in ice from the 40k world are remarkable in that they lack values higher than +0.5‰ and the total range of δ<sup>18</sup>O<sub>atm</sub> values is only about 20% of that observed in ice from the 100k world. Smaller changes in seawater δ<sup>18</sup>O values (Fig. 2) help to explain the absence of δ<sup>18</sup>O<sub>atm</sub> values higher than +0.5‰. However, additional mechanisms are likely to be needed to explain the lack of precession-modulated variability in δ<sup>18</sup>O<sub>atm</sub> associated with biosphere productivity and changes in the low-latitude water cycle<sup>21</sup>.

Plots of atmospheric CO<sub>2</sub> against δD<sub>ice</sub> and CH<sub>4</sub> against δD<sub>ice</sub> (Fig. 4) show that samples from the MPT and 40k world fall within the envelope of the 100k world but exhibit only a fraction of its range. Samples from the MPT and 40k world do not have the isotopically light δD<sub>ice</sub> and low CO<sub>2</sub> values characteristic of glacial samples in the 100k world. Both CO<sub>2</sub> and CH<sub>4</sub> are positively correlated with δD<sub>ice</sub> in ice from the 100k world (S27). By contrast, in the MPT and 40k world, only δD<sub>ice</sub> and CO<sub>2</sub> are significantly correlated ( $P = 0.002$  and  $P = 0.003$ , respectively). For all three sample sets, the slopes of the correlation between CO<sub>2</sub> and δD<sub>ice</sub> are statistically indistinguishable (Fig. 4), in agreement with previous reconstructions from boron-based proxies and benthic δ<sup>18</sup>O values. Together, these observations suggest a consistent and persistent coupling between Antarctic climate and atmospheric CO<sub>2</sub> throughout the Pleistocene<sup>8,9</sup>, and that glacial cycles in the 40k world are truncated versions of glacial cycles in the 100k world.



**Fig. 4 | Cross-correlations between climate properties.** **a**,  $\text{CO}_2$  versus  $\delta\text{D}_{\text{ice}}$ . **b**,  $\text{CH}_4$  versus  $\delta\text{D}_{\text{ice}}$ . Data points are colour-coded by age: Black, Vostok  $\text{CO}_2$  or  $\text{CH}_4$  data<sup>3</sup> projected onto coeval Allan Hills  $\delta\text{D}_{\text{ice}}$  from S27; orange, Allan Hills MPT data; red, Allan Hills 40k world data. Shading represents the range of  $\delta\text{D}_{\text{ice}}$  we observed in the 40k world.  $\delta\text{D}_{\text{ice}}$  measurements in samples measured for  $\text{CO}_2$  span 94% of the total range of  $\delta\text{D}_{\text{ice}}$  observed in 40k world samples, and therefore the  $\text{CO}_2$  samples represent nearly the full range of climate variability recorded in the cores. There are significant correlations between  $\text{CO}_2$  and  $\delta\text{D}_{\text{ice}}$  in all three age units, and their regression slopes (ppm by ‰; solid lines) are statistically indistinguishable:  $1.33 \pm 0.17$  ( $2\sigma$ ;  $r = 0.78$ ),  $1.14 \pm 0.68$  ( $r = 0.61$ ), and  $0.90 \pm 0.56$  ( $r = 0.48$ ) for 120–250 ka, MPT, and pre-MPT age units, respectively. For comparison, a significant ( $P < 0.05$ ) correlation between  $\text{CH}_4$  and  $\delta\text{D}_{\text{ice}}$  exists only at S27 ( $3.56 \pm 0.65$ ;  $r = 0.61$ ).

## Conclusions

Shallow ice cores drilled in the Allan Hills BIA provide a direct observation of the variability in atmospheric  $\text{CO}_2$  during the 40k world, confirming previous findings that minimum  $\text{CO}_2$  concentrations declined after the MPT<sup>8,9</sup>. These results complement plans to drill a continuous ice core record back to around 1.5 Ma<sup>27</sup>. These snapshots of the Earth's climate system from shallow ice cores in BIAs provide an incomplete picture, especially with regards to the dynamics. Nevertheless, the oldest ice will probably be found in discontinuous sections. Our work demonstrates

that BIAs can be exploited to extend climate records, including atmospheric  $\text{CO}_2$  concentrations, well back into the early Pleistocene and perhaps even the Pliocene.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1692-3>.

- Imbrie, J. et al. On the structure and origin of major glaciation cycles: 2. The 100,000-year cycle. *Paleoceanography* **8**, 699–735 (1993).
- Lisiecki, L. E. & Raymo, M. E. A Pliocene–Pleistocene stack of 57 globally distributed benthic  $\delta^{18}\text{O}$  records. *Paleoceanography* **20**, PA1003 (2005).
- Petit, J. R. et al. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).
- Jouzel, J. et al. Orbital and millennial Antarctic climate variability over the past 800,000 years. *Science* **317**, 793–796 (2007).
- Siegenthaler, U. et al. Stable carbon cycle–climate relationship during the late Pleistocene. *Science* **310**, 1313–1317 (2005).
- Lüthi, D. et al. High-resolution carbon dioxide concentration record 650,000–800,000 years before present. *Nature* **453**, 379–382 (2008).
- Imbrie, J. et al. On the structure and origin of major glaciation cycles: 1. Linear responses to Milankovitch forcing. *Paleoceanography* **7**, 701–738 (1992).
- Hönisch, B. et al. Atmospheric carbon dioxide concentration across the mid-Pleistocene transition. *Science* **324**, 1551–1554 (2009).
- Chalk, T. B. et al. Causes of ice age intensification across the mid-Pleistocene transition. *Proc. Natl Acad. Sci. USA* **114**, 13114–13119 (2017).
- Dyez, K. A. et al. Early Pleistocene obliquity-scale  $\text{pCO}_2$  variability at ~1.5 million years ago. *Paleoceanography* **33**, 1270–1291 (2018).
- Zachos, J. et al. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693 (2001).
- Clark, P. U. et al. The middle Pleistocene transition: characteristics, mechanisms, and implications for long-term changes in atmospheric  $\text{pCO}_2$ . *Quat. Sci. Rev.* **25**, 3150–3184 (2006).
- Köhler, P. & Bintanja, R. The carbon cycle during the mid Pleistocene transition: the southern ocean decoupling hypothesis. *Clim. Past* **4**, 311–332 (2008).
- Lisiecki, L. E. A benthic  $\delta^{13}\text{C}$ -based proxy for atmospheric  $\text{pCO}_2$  over the last 1.5 Myr. *Geophys. Res. Lett.* **37**, L21708 (2010).
- Higgins, J. A. et al. Atmospheric composition 1 million years ago from blue ice in the Allan Hills, Antarctica. *Proc. Natl Acad. Sci. USA* **112**, 6887–6891 (2015).
- Whillans, I. M. & Cassidy, W. A. Catch a falling star—meteorites and old ice. *Science* **222**, 55–57 (1983).
- Bender, M. L. et al. The contemporary degassing rate of Ar-40 from the solid Earth. *Proc. Natl Acad. Sci. USA* **105**, 8232–8237 (2008).
- Spaulding, N. E. et al. Climate archives from 90 to 250 ka in horizontal and vertical ice cores from the Allan Hills Blue Ice Area, Antarctica. *Quat. Res.* **80**, 562–574 (2013).
- Bereiter, B. et al. Diffusive equilibration of  $\text{N}_2$ ,  $\text{O}_2$  and  $\text{CO}_2$  mixing ratios in a 1.5-million-years-old ice core. *Cryosphere* **8**, 245–256 (2014).
- Pol, K. et al. New MIS 19 EPICA Dome C high resolution deuterium data: hints for a problematic preservation of climate variability at sub-millennial scale in the 'oldest ice'. *Earth Planet. Sci. Lett.* **298**, 95–103 (2010).
- Landais, A. et al. What drives the millennial and orbital variations of  $\delta^{18}\text{O}_{\text{atm}}$ ? *Quat. Sci. Rev.* **29**, 235–246 (2010).
- Raymo, M. E. et al. Influence of late Cenozoic mountain building on ocean geochemical cycles. *Geology* **16**, 649–653 (1988).
- Berger, A. et al. Modelling northern hemisphere ice volume over the last 3 Ma. *Quat. Sci. Rev.* **18**, 1–11 (1999).
- Martínez-García, A. et al. Southern ocean dust–climate coupling over the past four million years. *Nature* **476**, 312–315 (2011).
- Pena, L. D. & Goldstein, S. L. Thermohaline circulation crisis and impacts during the mid-Pleistocene transition. *Science* **345**, 318–322 (2014).
- Loulergue, L. et al. Orbital and millennial-scale features of atmospheric  $\text{CH}_4$  over the past 800,000 years. *Nature* **453**, 383–386 (2008).
- Fischer, H. et al. Where to find 1.5 million yr old ice for the IPICS 'oldest-ice' ice core. *Clim. Past* **9**, 2489–2505 (2013).
- Bereiter, B. et al. Revision of the EPICA Dome C  $\text{CO}_2$  record from 800 to 600 kyr before present. *Geophys. Res. Lett.* **42**, 542–549 (2015).
- Suwa, M. & Bender, M. L. Chronology of the Vostok ice core constrained by  $\text{O}_2/\text{N}_2$  ratios of occluded air, and its implication for the Vostok climate records. *Quat. Sci. Rev.* **27**, 1093–1106 (2008).
- Dreyfus, G. B. et al. Anomalous flow below 2700 m in the EPICA Dome C ice core detected using  $\delta^{18}\text{O}$  of atmospheric oxygen measurements. *Clim. Past* **3**, 341–353 (2007).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### Sample location and description

**Allan Hills Blue Ice Area (BIA).** The Allan Hills BIA is located ~100 km to the northwest of the McMurdo Dry Valleys. Drilling sites are located to the west of the Allan Hills nunatak in the Main Ice Field (MIF) of the Allan Hills BIAs (Extended Data Fig. 1). Here, net ablation leads to the exposure of ancient glacial ice at the surface of an ice sheet<sup>31</sup>. Elevation, bedrock topography, and ice velocities in the MIF have been previously documented<sup>32,33</sup>. Site meteorology and surface snow properties are discussed in more detail by Dadic et al.<sup>34</sup> and ice and gas properties described in Higgins et al.<sup>15</sup>.

**Site ALHIC1503 (previously named BIT-58).** Site ALHIC1503 (76.73243° S, 159.3562° E) is located near a crest of a northwest–southeast trending ice ridge off the MIF. The precise direction and the velocity of ice flow are not known at this location. Measurements from the nearby ice field are consistent with primary flow to the east or northeast. Complex dust bands are visible from high-resolution satellite imagery (Extended Data Fig. 1, left), implying a complex history of glacial flow. Typical ice flow velocities near ALHIC1503 are low (0.015 m yr<sup>-1</sup>)<sup>33</sup>.

Bedrock topography, determined using ground penetrating radar (GPR), indicates that ALHIC1503 is located on a steep (~45°) slope leading to a local bedrock high with an overlying ice thickness of ~150 m (Extended Data Fig. 2). In the 2015–2016 field season, 23 m of ice was drilled, extending the earlier 124-m-long ice core<sup>15</sup> to the bedrock at 147 m below the ice surface.

**Site ALHIC1502.** Site ALHIC1502 (76.73286° S, 159.35507° E) is located 56 m upflow from ALHIC1503. The ice thickness is ~204 m, consistent with a steep bedrock slope (Extended Data Fig. 2). Drilling at site ALHIC1502 recovered 197 m of ice.

**Site 27 (S27).** S27 (76.70° S, 159.31° E) is located along the main ice flow line of the Allan Hills region. Two hundred and twenty-four metres of ice was retrieved, representing ~70% of the estimated column thickness. Spaulding et al.<sup>18</sup> provide a comprehensive suite of stable water isotope and gas analyses for S27. The S27 ice chronology has been established by matching  $\delta D_{ice}$  features to those of the Vostok ice core. This timescale was validated by correlating variations in  $\delta^{18}O_{atm}$  ( $\delta^{18}O$  of paleoatmospheric  $O_2$ ) into the record of  $\delta^{18}O_{atm}$  in Vostok ice cores<sup>35</sup>.

### Analytical procedures

**$^{40}Ar_{atm}$  and  $\delta Xe/Kr$ .**  $^{40}Ar$  is produced in the solid earth by the radioactive decay of  $^{40}K$ . As  $^{40}Ar$  slowly leaks into the atmosphere, its concentration increases with time.  $^{38}Ar$  and  $^{36}Ar$ , on the other hand, are stable, primordial, and non-radiogenic, so their atmospheric concentrations are treated as constant. The ratio of  $^{40}Ar/^{38}Ar$  thus rises towards the future and decreases towards the past. The term of merit here is the gravitationally corrected paleo-atmospheric  $^{40}Ar/^{38}Ar$  ratio:  $^{40}Ar_{atm} = \delta^{40}Ar/^{38}Ar - \delta^{38}Ar/^{36}Ar$ .

The latter term corrects for the gravitational fractionation in the firn column<sup>36</sup>. Studies of  $^{40}Ar_{atm}$  as a function of age in the Dome C and Vostok ice cores constrained the rate of change to be  $0.066 \pm 0.006\%$  Myr<sup>-1</sup>, allowing us to date the air trapped in the ice without the prerequisite for stratigraphic continuity<sup>17</sup>.

A potential complication of  $^{40}Ar_{atm}$  dating comes from the in situ production of  $^{40}Ar$  from the bedrock and/or from dust, a tiny amount of which is present in polar ice cores. The deepest sample at ALHIC1502 has a much younger age (0.8 Myr) than the overlying ice (2.2 Myr). This inverted age–depth relationship may result either from folding or from radiogenic input of  $^{40}Ar$  from the  $^{40}K$  in the underlying bedrock, a phenomenon observed for basal ice at GISP2 in Greenland<sup>37</sup>. The  $^{40}K$  from dust in the ice would produce a negligible amount of  $^{40}Ar$ .

Ar isotope analyses on the original BIT-58 core between the surface and 126-m depth were carried out at Princeton University with a procedure modified from Bender et al.<sup>17</sup> and described by Higgins et al.<sup>15</sup>. The

standard deviation of  $^{40}Ar_{atm}$  measured across 68 external standards is  $\pm 0.014\%$ , corresponding to an age uncertainty of  $\pm 210$  kyr ( $1\sigma$ ). For Allan Hills samples drilled in 2015–2016,  $^{40}Ar_{atm}$  was measured at Scripps Institution of Oceanography with slight modification. First, sample size was increased from 500 to 800 g for better precision. Second, pumping time for evacuating the ice-bearing vessel before gas extraction was reduced to 15 min to prevent gas loss. Third, extracted and purified gases were measured on a newer mass spectrometer, which offers better signal sensitivity and stability. Finally, following published procedures<sup>38</sup>, we measured the Xe/Kr ratios, expressed as  $\delta Xe/Kr$ , an indicator of mean ocean temperature<sup>39</sup>, in the same aliquot of the extracted and getterred gases. A typical  $^{40}Ar_{atm}$  and  $\delta Xe/Kr$  measurement takes ~3.5 h on the mass spectrometer.

Two measurement campaigns at Scripps Institution of Oceanography were carried out and each was associated with slightly different analytical procedures. In the first campaign, a fixed amount of reference gas was introduced into the bellow of the mass spectrometer, leaving different amount of gas in sample and reference sides. As the sample and reference gases deplete during analysis, the pressure and ion current fall more rapidly for the side containing less gas. As a result,  $^{40}Ar_{atm}$  was observed to depend on the volume of the analyte relative to the reference. All measured  $^{40}Ar_{atm}$  values were subsequently corrected for volume differences, using the voltage readings of the mass-40 ( $^{40}Ar$ ) beam when the gases were introduced to the fully expanded bellow on the mass spectrometer. The magnitude of this correction is generally less than 0.02‰, or 300 kyr, and the uncertainty in the correction is much smaller. In the second campaign, the amount of gas in the standard-side bellows was matched to that of the reference-side bellows. No volume correction was applied in the second campaign. Based on replicate analyses, the external reproducibility was  $\pm 0.006\%$  ( $1\sigma$ ) and  $\pm 0.007\%$  ( $1\sigma$ ) for dry La Jolla air samples and Taylor Glacier blue-ice samples, respectively. The samples measured at Scripps therefore have an age uncertainty of  $\pm 110$  kyr ( $1\sigma$ ) owing to the analytical uncertainty, or 10% of the sample age owing to the uncertainty in the  $^{40}Ar$  outgassing rate, whichever is greater. Individual sample uncertainties are reported along with the measured values.

Xenon-to-krypton ratios ( $\delta Xe/Kr$ ) were measured immediately after the  $^{40}Ar_{atm}$  measurements on the same mass spectrometer by ‘peak hopping’ between  $^{84}Kr$  and  $^{132}Xe$ . The  $\delta Xe/Kr$  measurement takes about half an hour to complete. The final reproducibility is  $\pm 0.21\%$  and  $\pm 0.27\%$  for dry La Jolla air samples and Taylor Glacier ice replicates, respectively. In order to calculate mean ocean temperature from  $\delta Xe/Kr$ , which is normalized to La Jolla air, we need to make an assumption about the volume of the ocean<sup>39</sup>. For this purpose, we use a sensitivity of  $1.1 \pm 0.2$  °C/‰. We obtain this value from the observed last glacial maximum (LGM)–present  $\delta Xe/Kr$  change<sup>39</sup> with minor modifications. We adopted all the same sea-level-related corrections as in ref.<sup>39</sup> except one: the assumption that Kr and Xe saturation state was higher during the LGM than at present. This assumption lowers the sensitivity of mean ocean temperature (MOT) to  $\delta Xe/Kr$  from 1.2 to 1.0 °C/‰ (for example, a MOT change of  $2.57 \pm 0.24$  °C for a 2.5‰ change of  $\delta Xe/Kr$ <sup>39</sup>). To account for this uncertainty we adopt an intermediate value of 1.1 and a larger error of  $\pm 0.2$ .

**$\delta D_{ice}$  and  $\delta^{18}O_{ice}$ .** Stable water isotope measurements ( $\delta^{18}O_{ice}$ ,  $\delta D_{ice}$ ) were taken from a slab of ice, 2.5-cm wide, cut (parallel to the vertical axis) from the outside of each ice core. Because some portions of the core were fractured, continuous sections were not available at all depths. These slabs were sub-sampled at 10–15-cm resolution at the Climate Change Institute, University of Maine. Each sub-sample was melted in a sealed plastic bag at room temperature, vigorously shaken, and then decanted into an 18-ml plastic scintillation vial. All vials were refrozen and stored below –10 °C until the time of analysis. Water isotope samples were measured via cavity ring-down spectroscopy (CRDS) using a Picarro Model L2130-i Ultra High-Precision Isotopic Water Analyzer coupled with a High Precision Vaporizer and liquid auto sampler module.  $\delta^{18}O_{ice}$  and  $\delta D_{ice}$  were measured simultaneously with an internal precision ( $2\sigma$ )

of  $\pm 0.05\%$  for  $\delta^{18}\text{O}_{\text{ice}}$ , and  $\pm 0.10\%$  for  $\delta\text{D}_{\text{ice}}$ . The instrument drift adjustment and daily calibration were accomplished by periodically measuring three secondary standards: BBB (an average Maine freshwater from the Bear Brook watershed), ASS (Antarctic surface snow) and LAP (light Antarctic precipitation). These standards were calibrated against the internationally accepted water isotope standards: SMOW (standard mean ocean water), SLAP (standard light Antarctic precipitation) and GISP (Greenland ice sheet precipitation).

**$\delta^{15}\text{N}$ ,  $\delta^{18}\text{O}_{\text{atm}}$ , and  $\delta\text{O}_2/\text{N}_2$ .** Analyses for the original BIT-58  $\text{O}_2/\text{N}_2/\text{Ar}$  elemental ratio,  $\delta^{18}\text{O}$  of  $\text{O}_2$ , and  $\delta^{15}\text{N}$  of  $\text{N}_2$  were carried out as described<sup>15,30</sup>. Samples from ALHIC1502 and ALHIC1503 were measured using a slightly modified procedure. In brief, ~20 g of ice was placed into a glass flask chilled in a dry-ice-isopropanol cold bath ( $-78.5^\circ\text{C}$ ) and the vessel was turbo-pumped for 3 min. The final pressure of the residual gas inside the vial was no higher than 1.5 mTorr. The pressure of the air trapped in ice, when expanded to the same space, was ~3 Torr. The reduced pumping time thus introduces, at most, about 0.05% of modern atmosphere into the sample gas.

After melting, the gas–meltwater mixture was equilibrated for at least 4 h. The majority of the meltwater was drained and leftover water inside the glass vial refrozen at  $-30^\circ\text{C}$ . Once all meltwater had refrozen, headspace gases were collected by condensation at liquid helium temperatures. The sample was admitted to a mass spectrometer (Thermo Finnegan Delta Plus XP) for elemental and isotope ratio analysis after homogenizing at room temperature for 30 min. The reproducibility of  $\delta^{15}\text{N}$ ,  $\delta\text{O}_2/\text{N}_2$  and  $\delta^{18}\text{O}_{\text{atm}}$  values from ALHIC1502 and ALHIC1503 samples analysed using this procedure, calculated as pooled standard deviation, is  $\pm 0.008\%$ ,  $\pm 3.164\%$ , and  $\pm 0.024\%$ , respectively.

Overall, the reduced pumping time leads to a reduction in gas loss and an improvement in precision. We note that the gas loss processes that alter  $\delta\text{O}_2/\text{N}_2$  did not appear to affect the isotopic composition of  $\text{O}_2$  and  $\text{N}_2$  in our samples (Extended Data Fig. 3). As a result, no gas loss correction was applied to the gas isotopes, in contrast to some previous studies in heavily fractured ice (for example, Siple Dome<sup>40</sup>). All  $\delta^{15}\text{N}$ ,  $\delta\text{O}_2/\text{N}_2$  and  $\delta^{18}\text{O}_{\text{atm}}$  values were normalized to modern atmosphere.

**$\text{CH}_4$ ,  $\text{CO}_2$  and  $\delta^{13}\text{C}$  of  $\text{CO}_2$ .**  $\text{CH}_4$  concentrations were analysed at Oregon State University using a melt refreeze technique<sup>41</sup>. Samples (~60–70 g of ice) were trimmed, melted under vacuum, and refrozen at about  $-70^\circ\text{C}$ .  $\text{CH}_4$  concentrations in released air were measured by gas chromatography and referenced to air standards calibrated by National Oceanic and Atmospheric Administration Global Monitoring Division (NOAA GMD) on the NOAA04 scale. Precision was generally better than  $\pm 4$  ppb. Individual sample uncertainties, where available, are reported along with the measured values.

$\text{CO}_2$  concentrations, referenced to air standards calibrated by NOAA GMD on the World Meteorological Organization (WMO) scale, were measured using a dry extraction (crushing) method<sup>42</sup>. Wherever possible, ice samples were processed and analysed in replicate for each depth and results averaged to obtain final  $\text{CO}_2$  concentrations. However, four pairs of replicates did not come from exactly the same depth and their reproducibility showed substantial deterioration beyond usual analytical uncertainties. In this case, we treated those four pairs of replicates as eight individual, unreplicated data points. For greenhouse gas concentrations, no gravitational fractionation correction was applied; the correction would be less than 0.8 ppb for  $\text{CH}_4$  and 0.5 ppm for  $\text{CO}_2$ .

The measurement protocol for  $\delta^{13}\text{C}$  of  $\text{CO}_2$  was as described<sup>43</sup> with improvements in the efficiency of air extraction. Approximately 200 g of ice was dry-extracted with a custom-made ‘ice grater’. The reproducibility of the method, based on replicate analysis of Taylor Glacier blue-ice samples, is  $0.02\%$  for  $\delta^{13}\text{C}\text{-CO}_2$ . The method also provides a measurement of  $\text{CO}_2$  concentrations. The uncertainties are  $\pm 2$  ppm (1 $\sigma$  of the pooled standard deviation). The final  $\delta^{13}\text{C}$  values, normalized to the Vienna–Pee Dee Belemnite (VPDB) standard, were corrected for blanks,  $\text{N}_2\text{O}^+$  mass interference from  $\text{N}_2\text{O}$ , and gravitational fractionation using the nearest  $\delta^{15}\text{N}$  value.

## Age assignment of $\text{CO}_2/\text{CH}_4$ and $\text{O}_2/\text{N}_2/\text{Ar}$ samples

The depths of samples analysed for  $\text{CO}_2/\text{CH}_4$  and  $\text{O}_2/\text{N}_2/\text{Ar}$  were different from depths of the  $^{40}\text{Ar}_{\text{atm}}$  samples that provided the chronology.  $\text{CO}_2/\text{CH}_4$  samples,  $\text{O}_2/\text{N}_2/\text{Ar}$  samples, and ice samples were assigned ages by bracketing  $^{40}\text{Ar}_{\text{atm}}$  values. Here we discuss two different models for assigning ages and show that different criteria for age assignment lead to similar conclusions. For the purpose of investigating the relationship between greenhouse gases ( $\text{CO}_2$  and  $\text{CH}_4$ ) and Antarctic climate ( $\delta\text{D}_{\text{ice}}$ ), we binned samples into three units: MPT, 800–1,200 ka; pre-MPT, >1,200 ka; and young ice, <800 ka. We focus on the robustness of ages for  $\text{CO}_2$  and  $\text{CH}_4$  samples. All  $\text{CO}_2$  and  $\text{CH}_4$  data discussed below exclude samples affected by respiration.

**Conservative approach.** In the most conservative approach, an age was assigned only to samples bracketed by two  $^{40}\text{Ar}_{\text{atm}}$  ages lying within a single age bin. Otherwise, the sample was left ‘unassigned’. In this binning scenario, 19  $\text{CO}_2$  samples were classified as MPT, 15 as pre-MPT, and 34 as unassigned. We classified 28 and 19  $\text{CH}_4$  samples as MPT and unassigned, respectively. Eight  $\text{CH}_4$  samples were binned into pre-MPT.

**Proximity approach.** In this approach, each  $\text{CO}_2/\text{CH}_4$  datum was assigned the  $^{40}\text{Ar}_{\text{atm}}$  age of the closest Ar isotope sample. In this age model, 23 and 37  $\text{CO}_2$  samples were binned into the MPT and pre-MPT units, respectively. For  $\text{CH}_4$ , 31 samples were classified as MPT, and 16 samples as pre-MPT. Finally, eight  $\text{CO}_2$  and eight  $\text{CH}_4$  samples were binned into the younger (<800 ka) age unit.

Extended Data Fig. 4 summarizes the  $\text{CO}_2\text{--}\delta\text{D}_{\text{ice}}$  and  $\text{CH}_4\text{--}\delta\text{D}_{\text{ice}}$  relationships using the two age models. Our conclusions regarding the greenhouse gas– $\delta\text{D}_{\text{ice}}$  relationship are not dependent on which model is used. As the proximity approach assigns an age to every depth, we adopted it for our interpretations.

## Assessing the fraction of climate variability recovered by discrete sampling

When interpreting data from stratigraphically disturbed ice, the question arises as to what fraction of the amplitude of climate variability is accessed in the suite of samples analysed. Here we consider two factors: (1) the fraction of climate variability preserved in the ice cores compared to the true natural variability; and (2) to what extent discrete samples capture the variability that is preserved in the ice. Below we discuss these two factors separately.

**Fraction of the pre-MPT  $\delta\text{D}_{\text{ice}}$  range preserved in the ice.** We start by considering what fraction of the true glacial–interglacial variability in the 40k world is preserved in the ice. If we assume the absence of post-depositional alteration processes such as respiration and diffusive smoothing, this fraction should be an intrinsic property of the ice itself and insensitive to the property being measured. Given the large number (>500) and the high spatial resolution (10 cm) of  $\delta\text{D}_{\text{ice}}$  measurements, we used  $\delta\text{D}_{\text{ice}}$  to evaluate the fraction of true  $\delta\text{D}_{\text{ice}}$  variability preserved in the ice. Next, we needed to make an assumption of the true  $\delta\text{D}_{\text{ice}}$  amplitude in the 40k world. It is assumed that the true amplitude of  $\delta\text{D}_{\text{ice}}$  variations in Allan Hills ice scales linearly with the amplitude of deep ocean temperature. This assumption is justified for two reasons. First, most of the bottom water formation today occurs in the southern ocean. Second, Antarctic temperature has been found to correlate tightly with mean ocean temperature during the LGM–Holocene transition<sup>39</sup> and with South Pacific bottom water temperature through the past 800 kyr<sup>44</sup>.

For deep ocean temperature, we considered two scenarios. First, we assumed a constant partitioning between ocean temperature and the influence of seawater  $\delta^{18}\text{O}$  on the benthic foram  $\delta^{18}\text{O}$  record throughout the Pleistocene. This means that the amplitude of  $\delta\text{D}_{\text{ice}}$  should simply scale with the benthic foram  $\delta^{18}\text{O}$  values. The amplitudes of glacial cycles in the early and mid-Pleistocene (between marine isotope stages (MIS) 39 and 104), expressed in benthic foram  $\delta^{18}\text{O}$ , averages  $0.70\%$ , representing 37% of the change from MIS 6 to MIS 5e ( $1.90\%$ ).



The range of  $\delta D_{ice}$  values between MIS 6 and MIS 5e observed in the Allan Hills ice from the 100k world was 80‰. Multiplying this 100k-world range by the scaling factor (37%) based on benthic foram  $\delta^{18}O$  predicts the range of  $\delta D_{ice}$  in 40k-world ice to be 30‰. By comparison, the range of 40k-world  $\delta D_{ice}$  observed in the Allan Hills ice is 32‰. Using this approach, we estimated that our pre-MPT ice samples represent ~100% of the average amplitude of glacial–interglacial  $\delta D_{ice}$  cycles in the 40k world.

In an alternative scenario, we considered deep ocean temperature inferred from a seawater  $\delta^{18}O$  record constrained independently from records of carbonate microfossil  $\delta^{18}O$  from the Mediterranean Sea<sup>45</sup>. A slightly higher variability in the 40k world deep ocean temperature, which corresponds to 50% of the average variability in the 100k world, lowers our estimate of the recovered climate variability to ~80%. Consequently, although other factors might influence the recovery of the range of climate variability, given these two approaches our best estimate is that the Allan Hills ice samples presented here preserve 90% of the true range of the climate variables in the 40k world.

**Fraction of the pre-MPT CO<sub>2</sub> and CH<sub>4</sub> range recovered by discrete samples.** CO<sub>2</sub> and CH<sub>4</sub> concentrations were measured on a relatively small number of samples. It is therefore possible that we have not retrieved the full range of variability of these properties in the 40k world, even if their true variability is well preserved in the ice cores. We evaluated this possibility by first generating a synthetic CO<sub>2</sub> series based upon the assumption that the variation of CO<sub>2</sub> in the 40k world scales linearly with benthic foram  $\delta^{18}O$  values. A synthetic CO<sub>2</sub> time series between 1.2 and 2.0 Ma was constructed on the basis of the LR04 global benthic foram  $\delta^{18}O$  stack<sup>2</sup>. A linear regression between benthic  $\delta^{18}O$  and atmospheric CO<sub>2</sub> between 0 and 800 ka resulted in correlation parameters that were used to create an artificial atmospheric CO<sub>2</sub> record further back in time (Extended Data Fig. 5a, b). After resampling to interpolate the synthetic data at a temporal resolution of 2.5 kyr, a Monte Carlo simulation randomly selected 37 samples from the synthetic CO<sub>2</sub> record. The simulation was repeated 10,000 times. We then calculated the observed range of the 37 random CO<sub>2</sub> samples and compared it to the ‘true’ range, yielding a ratio for each simulation run. Finally, the distribution of this ratio is shown as a series of histograms (Extended Data Fig. 5c–h). Note that the term of merit for each simulation is the ratio of the sampled range to the hypothetical true range.

To account for the fact that interglacial accumulation rates are generally higher than glacial rates<sup>46</sup>, we multiplied the modelled occurrence of interglacial ice (CO<sub>2</sub> > 250 ppm) by 2, 4, 8, 16, and 32. When the presence of the interglacial ice is two times the glacial ice,  $84 \pm 15\%$  (95% confidence interval) of the total CO<sub>2</sub> range can be captured by 37 samples (Extended Data Fig. 5d). This ratio was derived from Vostok, an inland coring site of East Antarctica<sup>47</sup>.

We note that 35 out of the 37 40k-world CO<sub>2</sub> samples involve replicates measured at the same depth. The assumption here is that replicates at the same depth have exactly the same age, which remains unevaluated in Allan Hills ice cores. It is possible that the ice flow and dip of ice layers make samples from the same depth different in age. If this is the case, each pair of replicates should be treated as two individual measurements, and the total number of samples would be 72. In this scenario, the recovered range is from 214 to 281 ppm, and 72 discrete samples capture  $88 \pm 13\%$  of the CO<sub>2</sub> variability preserved in the ice. Treating replicates as individual samples does not significantly change the range (from 65 to 67 ppm), but it increases the likelihood that we are recovering close to the full range of CO<sub>2</sub> in the ice (from 84% to 88%).

Considering the discussion above, our best estimate is that the 37 CO<sub>2</sub> samples cover  $76 \pm 14\%$  of the true CO<sub>2</sub> range in the 40k world (90% of true variability preserved in the ice multiplied by  $84 \pm 15\%$  that is captured by discrete sampling). Furthermore, we note that the  $\delta D_{ice}$  values of the ice samples analysed for CO<sub>2</sub> as well as CH<sub>4</sub> span more than 90% of the observed 40k-world  $\delta D_{ice}$  range (Fig. 4). This number is higher than the estimate yielded by Monte Carlo simulations, meaning that the discrete samples are likely to capture more variability than probability alone would dictate. As a result, there is a greater chance that we have

captured a significant fraction (>90%) of the glacial–interglacial CO<sub>2</sub> range in the 40k world.

For  $\delta^{18}O_{atm}$  values the estimated recoverability should be no less than 70%. First, 29 samples were analysed for  $\delta^{18}O_{atm}$ . Second, the co-depth  $\delta D_{ice}$  values of  $\delta^{18}O_{atm}$  samples span 97% of the 40k-world  $\delta D_{ice}$  (Extended Data Fig. 6a).

### Potential sample mixing on various length scales

**Mixing by molecular diffusion.** The interval between 123 and 141 m in ALHIC1503 shows unexpectedly low variability in the elemental and isotopic ratios of the major gases (Extended Data Fig. 7). Molecular diffusion in the firn column has been shown to reduce annual variability of water isotopes<sup>48</sup> and gases<sup>49</sup>. In polycrystalline ice, diffusive mixing has also been invoked to explain the lack of sub-millennial  $\delta D_{ice}$  variability preserved in the EPICA Dome C ice core, dated ~780 kyr old<sup>20</sup>. When annual layers are thin and temperatures are high, some gas records (for example,  $\delta O_2/N_2$ ) can also undergo extensive diffusive smoothing and are expected to lose even the glacial–interglacial variability<sup>20</sup>.

However, diffusive smoothing of climate records is unlikely to be significant at the Allan Hills BIA. First, current temperatures within the boreholes at the Allan Hills BIA are around ~30 °C<sup>32</sup>. Rates of diffusion at the Allan Hills BIA should be substantially slower than at the warm basal temperatures (>–10 °C) expected at the bottom of thick (>3 km) polar ice sheets. Second, as is shown below, water isotopes ( $\delta D_{ice}$ ) are expected to be much more sensitive to diffusive smoothing than gases (O<sub>2</sub> and N<sub>2</sub>). As a result, the presence of substantial sample-to-sample variability in  $\delta D_{ice}$  values measured at 10-cm resolution indicates that the effects of diffusion on this and longer length scales are likely to be minor.

Below we quantitatively estimate the characteristic diffusion time scale of water isotopes and of oxygen and nitrogen gases in the ice. Gas and water permeation coefficients in the ice have been estimated from molecular dynamics simulations<sup>50,51</sup> and limited observations in some ‘natural experiments’<sup>52,53</sup>. Values obtained from different methods can differ by more than one order of magnitude. In our calculation, the fastest permeation parameters are used whenever possible to provide the most rigorous test of diffusive smoothing.

For the self-diffusion of water molecules in the ice, the diffusion time scale ( $\tau_{ice}$ ) depends solely on the diffusivity of water molecules in ice ( $D_{ice}$ ) and the length of interest ( $L$ ; 0.1 m):

$$\tau_{ice} = \frac{L^2}{D_{ice}} \quad (1)$$

The diffusivity of ice is determined by<sup>54</sup>:

$$D_{ice} = D_{ice}^0 \times \exp\left(-\frac{Q_{ice}}{RT}\right) \quad (2)$$

where  $D_{ice}^0$  ( $9.1 \times 10^{-4} \text{ m}^2 \text{ s}^{-1}$ ) is a constant of diffusion for water molecules in ice,  $Q_{ice}$  ( $5.86 \times 10^4 \text{ J mol}^{-1}$ ) the activation energy of diffusion,  $R$  ( $8.314 \text{ J mol}^{-1} \text{ K}^{-1}$ ) the gas constant, and  $T$  the temperature. The characteristic time scale of self-diffusion in ice as a function of temperature is shown in Extended Data Fig. 8. Depending on temperature, it takes  $10^5$  to  $10^7$  years for diffusive smoothing to smear the original signal on the length scale of 0.1 m.

Next, we consider the molecular diffusion of O<sub>2</sub> and N<sub>2</sub> in the ice. Below is a simple model of gas permeation with a large reservoir of gases (that is, bubbles) and a medium through which diffusion takes place (that is, ice): (1) First, gases in the reservoir dissolve into the ice. The amount of dissolved gas depends on ice volume, pressure, and solubility. (2) Next the gas will diffuse in the ice driven by the concentration gradient. This is assumed to be the rate-limiting step. (3) Diffusive gas exchange in the ice along the concentration gradient will immediately be compensated by the gas exchange between the ice phase and the bubble phase. (4) Eventually the concentration gradient in the bubbles will be balanced owing to diffusive flux.

# Article

Were there no gas reservoirs, the characteristic diffusion time scale of gas species  $m$  would simply follow the diffusivity of gas  $m$  in ice ( $D_m$ ) and the length of interest ( $L$ ):

$$\tau_m = \frac{L^2}{D_m} \quad (3)$$

where  $L$  is 0.5 m because the spatial resolution of gas measurements is about 50 cm.

Similar to the diffusivity of water isotopes, the diffusivity of gas  $m$  follows<sup>50</sup>:

$$D_m = D_m^0 \times \exp\left(-\frac{Q_{m,\text{ice}}}{RT}\right) \quad (4)$$

where  $D_m^0$  is a constant and  $Q_{m,\text{ice}}$  is the activation energy of diffusion.

However, the presence of the reservoir would compensate for the diffusive flux. Here, we introduce the concept of partitioning function  $Z$ , the ratio of the number of gas molecules in the bubbles to the number of gases dissolved in the ice.

$$\tau_m = Z_m \times \frac{L^2}{D_m} \quad (5)$$

By definition,  $Z$  of a given gas species  $m$  is

$$Z_m = \frac{n_{m,\text{gas}}}{n_{m,\text{ice}}} \quad (6)$$

where  $n_{m,\text{gas}}$  and  $n_{m,\text{ice}}$  are the number of gas molecules in the gas phase and in the ice phase, respectively.

It is assumed that  $n_{m,\text{gas}} \gg n_{m,\text{ice}}$ . The number of gas  $m$  molecules in the gas phase can be inferred from the molar fraction of gas  $m$  ( $f_m$ ) and the total number of gases in the ice ( $n_{\text{gas}}$ ), which has indeed been measured as the gas content of the ice ( $V$ ), expressed in SI units  $\text{m}^3 \text{kg}^{-1}$ :

$$n_{m,\text{gas}} = n_{\text{gas}} \times f_m \quad (7)$$

Given the ideal gas law, equation (7) can be re-written as:

$$n_{m,\text{gas}} = \frac{p^0 V m_{\text{ice}}}{RT^0} \times f_m \quad (8)$$

where  $m_{\text{ice}}$  is the mass of the ice,  $p^0$  is 101,325 Pa and  $T^0$  is 273 K.

The number of molecules of gas  $m$  in the ice phase depends on the solubility of gas  $m$  ( $S_m$ ), the partial pressure of gas  $m$  ( $P_m$ ), and the number of water molecules ( $n_{\text{ice}}$ ):

$$n_{m,\text{ice}} = S_m \times P_m \times n_{\text{ice}} \quad (9)$$

The solubility of gas  $m$  is governed by:

$$S_m = S_m^0 \times \exp\left(-\frac{Q_m}{RT}\right) \quad (10)$$

where  $S_m^0$  [ $\text{Pa}^{-1}$ ] is the dissolution constant, and  $Q_m$  the activation energy of dissolution.  $S_{\text{O}_2}^0$  and  $S_{\text{N}_2}^0$  are  $3.7 \times 10^{-13} \text{ Pa}^{-1}$  and  $4.5 \times 10^{-13} \text{ Pa}^{-1}$ , respectively;  $Q_{\text{O}_2}$  and  $Q_{\text{N}_2}$  are  $9,200 \text{ J mol}^{-1}$  and  $7,900 \text{ J mol}^{-1}$ , respectively<sup>50</sup>.

The partial pressure of gas  $m$  is the product of the bubble pressure, which is assumed to be the hydrostatic pressure of the depth in which bubbles are located, and the molar fraction of the gas  $m$  in air:

$$P_m = (\rho_{\text{ice}} \times g \times h) \times f_m \quad (11)$$

where  $\rho_{\text{ice}}$  is the density of ice ( $920 \text{ kg m}^{-3}$ ),  $g$  the gravitational acceleration constant ( $9.8 \text{ m s}^{-2}$ ), and  $h$  the depth. In our case,  $h$  is assumed to be a constant 150 m.

The number of water molecules  $n_{\text{ice}}$  can be directly calculated from the mass of the ice:

$$n_{\text{ice}} = \frac{m_{\text{ice}}}{M_{\text{water}}} \quad (12)$$

where  $M_{\text{water}}$  is the molecular weight of  $\text{H}_2\text{O}$  ( $0.018 \text{ kg mol}^{-1}$ ).

Therefore, the final expression for  $Z_m$  is:

$$Z_m = \frac{\frac{p^0 V m_{\text{ice}}}{RT^0} \times f_m}{S_m \times \rho_{\text{ice}} \times g \times h \times f_m \times \frac{m_{\text{ice}}}{M_{\text{water}}}} \quad (13)$$

$$Z_m = \frac{p^0 \times V \times M_{\text{water}}}{S_m \times \rho_{\text{ice}} \times g \times h \times R \times T^0} \quad (14)$$

Using equations (4) and (14) and relevant parameters, the diffusion time scale of  $\text{O}_2$  and  $\text{N}_2$  are computed as a function of absolute temperature (Extended Data Fig. 8). Unless at very cold temperatures ( $<230 \text{ K}$ ), water isotopes measured every 10 cm should be more susceptible to diffusive mixing than gases measured every 50 cm. As a result, the preservation of  $\delta\text{D}_{\text{ice}}$  variability argues that flat  $\delta\text{O}_2/\text{N}_2$  and  $\delta^{18}\text{O}_{\text{atm}}$  profiles are not the result of diffusive smoothing. However, if the temperature is indeed that low, the characteristic time scale of gases becomes too large ( $>10^7 \text{ yr}$ ) for substantial smoothing to occur in million-year-old ice. We did not calculate the permeation coefficients for  $\text{CH}_4$  or  $\text{CO}_2$  as these molecules have larger molecular diameters than  $\text{O}_2$  and  $\text{N}_2$ , and should diffuse at an even slower rate and have a longer  $\tau$  (ref. <sup>19</sup>).

**Mixing within individual samples.** In polar ice sheets the thickness of annual ice layers declines with depth owing to compression and lateral flow. Thinning will reduce geochemical variability if a property changes over timescales shorter than that encompassed by the physical length of a single sample. This problem is further amplified in stratigraphically disturbed ice, as the orientation of the layering is uncertain.  $\text{CO}_2$  and  $\text{O}_2/\text{N}_2/\text{Ar}$  samples were cut from smaller lengths of ice than  $\delta\text{D}_{\text{ice}}$  samples (which were 10-cm long). Therefore, assuming the layering is horizontal,  $\text{CO}_2$  and  $\text{O}_2/\text{N}_2/\text{Ar}$  samples should average shorter time intervals than is the case for  $\delta\text{D}_{\text{ice}}$  samples. On the other hand,  $\text{CH}_4$  samples were cut in 10-cm lengths. Because  $\delta\text{D}_{\text{ice}}$  samples show large variability, we consider that gas properties other than  $^{40}\text{Ar}_{\text{atm}}$  and  $\delta\text{Xe}/\text{Kr}$  represent relatively short climate intervals without much mixing. The  $\delta\text{D}_{\text{ice}}$  samples themselves may sometimes be mixed, and this possibility remains to be examined by the analysis of samples smaller than 10 cm.

**Flow-induced mixing.** Physical mixing of ice of different ages can result from glacial flow. Mechanical mixing of this type may create an artificial correlation between properties that are otherwise uncorrelated. Granted, many climate properties will co-vary in the absence of mixing. Therefore, property–property correlations do not necessarily indicate that mixing has occurred.

Extended Data Fig. 9a shows the Pearson correlation coefficient (expressed as  $R^2$ ) between  $\delta\text{D}_{\text{ice}}$  and  $d$  within intervals of varying lengths. In normally ordered ice cores,  $d$  is found to have a complex relationship with  $\delta\text{D}_{\text{ice}}$ <sup>55</sup>. We calculate  $R^2$  as a function of the number of  $\delta\text{D}_{\text{ice}}$  samples included. The section between 132 and 142 m shows a very strong correlation between  $\delta\text{D}_{\text{ice}}$  and  $d$ , which raises the suspicion of mixing.

We next examined the gas records between 132 and 140 m. The maximum and minimum  $\text{CO}_2$  values within this 10-m section are 253 and 217 ppm, respectively. Because the gas content in the deep ALHIC1502

and ALHIC1503 samples is very similar ( $\sim 0.098 \text{ cm}^3 \text{ g}^{-1}$ ), a simple linear mixing curve in the gas phase is expected. If we regard these two points as end members and plot  $\delta D_{\text{ice}}$  against  $\text{CO}_2$ , the data points do not fall onto a mixing line (Extended Data Fig. 9b). Similarly, it is difficult to explain the  $\delta D_{\text{ice}}-\text{CH}_4$  and  $\text{CO}_2-\text{CH}_4$  plots by simple two-end-member mixing. In other words, a simple mixing scenario with two end members cannot explain our gas data, measured at every  $\sim 50 \text{ cm}$ . We therefore consider the property–property plots (Fig. 4) to be unaffected by mechanical mixing on the length scale of  $50 \text{ cm}$ , although mixing at a much smaller length scale is still possible.

### Effect of respiration on gas properties

Respiration of detrital organic matter has been shown to lead to the in situ production of  $\text{CO}_2$  in the bottom of polar ice sheets<sup>56,57</sup>, resulting in elevated concentrations of greenhouse gases compared to the contemporaneous atmosphere. We observed substantially elevated  $\text{CO}_2$  and  $\text{CH}_4$  concentrations in the basal sections of ALHIC1503 and ALHIC1502, which we attribute to respiration and methanogenesis, respectively. Two lines of evidence support this hypothesis. First, there are extremely depleted ( $< -500\%$ )  $\delta \text{O}_2/\text{N}_2$  ratios and high ( $> 4\%$ )  $\delta^{18}\text{O}_{\text{atm}}$  values in samples near the bottom of ALHIC1503. For comparison, typical  $\delta \text{O}_2/\text{N}_2$  values in Antarctic ice cores vary between  $-5$  to  $-20\%$ <sup>58</sup> and  $\delta^{18}\text{O}_{\text{atm}}$  ranges from  $-0.5$  to  $+1.5\%$ <sup>21</sup> on glacial–interglacial timescales. Second, measured  $\delta^{13}\text{C}$  of  $\text{CO}_2$  values in some deep samples from the ALHIC1502 and ALHIC1503 cores (Extended Data Fig. 10) are outside the range ( $-6$  to  $-7\%$ ) observed for the last  $160 \text{ kyr}$ <sup>59</sup>. This is consistent with contributions from respiratory  $\text{CO}_2$  with a  $\delta^{13}\text{C}$  value of about  $-25\%$ . The deepest measured  $\delta^{13}\text{C}$  sample comes from  $3 \text{ m}$  above the bedrock in ALHIC1503 and has a  $\delta^{13}\text{C}$  value of  $-22.4\%$ . Assuming an initial  $\delta^{13}\text{C}$  of  $\text{CO}_2$  value of  $-6.5\%$  and an  $\delta^{13}\text{C}$  of  $\text{CO}_2$  value of  $-25\%$  from respired carbon, isotope mass balance suggests that  $86\%$  of the  $\text{CO}_2$  in this sample is derived from respiration.

In four out of the nine  $\delta^{13}\text{C}$  samples,  $\delta^{13}\text{C}$  of  $\text{CO}_2$  was compatible with the absence of respiration, assuming a cut-off  $\delta^{13}\text{C}$  value of  $-7\%$ . In the other five samples, which originated within  $7 \text{ m}$  of the bottom of the two cores, however,  $\delta^{13}\text{C}$  of  $\text{CO}_2$  was lighter than  $-7\%$  and indicates respiration. We marked the shallowest  $\delta^{13}\text{C}$  sample with less than  $-7\%$  isotopic as a cut-off depth and rejected all  $\text{CO}_2$  samples below it. Admittedly it is still possible that samples immediately above those cut-off depths might also be affected, but the highest pre-MPT  $\text{CO}_2$  value is found around  $131 \text{ m}$  in ALHIC1503, a depth that is bracketed by two  $\delta^{13}\text{C}$  measurements falling between  $-6\%$  and  $-7\%$  and is considered respiration-free. As a result, even if certain samples above the threshold depths are contaminated by respiration, our conclusions would not change. On the basis of these results, along with anomalous values of  $\text{CH}_4$ ,  $\delta \text{O}_2/\text{N}_2$ , and  $\delta^{18}\text{O}_{\text{atm}}$  in the deeper sections, we excluded biogenic gas data below  $185.950 \text{ m}$  in ALHIC1502 and  $139.625 \text{ m}$  in ALHIC1503 from our evaluation. We still included  $\delta D_{\text{ice}}$  from those sections in Fig. 2.

### Data availability

Allan Hills stable water isotope and gas data that support the findings of this study are available on the United States Antarctic Program Data Center (<http://www.usap-dc.org/>) with the following identifiers: DOI: 10.15784/601129 (ALHIC1502 stable water isotopes); DOI: 10.15784/601128 (ALHIC1503 stable water isotopes); DOI: 10.15784/601201 (heavy noble gases); DOI: 10.15784/601202 ( $\text{CO}_2$  concentration and  $\delta^{13}\text{C}-\text{CO}_2$ ); DOI: 10.15784/601203 ( $\text{CH}_4$  concentration); and DOI: 10.15784/601204 (elemental and isotopic composition of  $\text{O}_2$ ,  $\text{N}_2$  and Ar).

31. Bintanja, R. On the glaciological, meteorological, and climatological significance of Antarctic blue ice areas. *Rev. Geophys.* **37**, 337–359 (1999).
32. Delisle, G. & Sievers, J. Sub-ice topography and meteorite finds near the Allan Hills and the near Western Ice Field, Victoria Land, Antarctica. *J. Geophys. Res. Planets* **96**, 15577–15587 (1991).

33. Spaulding, N. E. et al. Ice motion and mass balance at the Allan Hills blue-ice area, Antarctica, with implications for paleoclimate reconstructions. *J. Glaciol.* **58**, 399–406 (2012).
34. Dacic, R. et al. Extreme snow metamorphism in the Allan Hills, Antarctica, as an analogue for glacial conditions with implications for stable isotope composition. *J. Glaciol.* **61**, 1171–1182 (2015).
35. Bender, M. L. Orbital tuning chronology for the Vostok climate record supported by trapped gas composition. *Earth Planet. Sci. Lett.* **204**, 275–289 (2002).
36. Craig, H. et al. Gravitational separation of gases and isotopes in polar ice caps. *Science* **242**, 1675–1678 (1988).
37. Bender, M. L. et al. On the nature of the dirty ice at the bottom of the GISP2 ice core. *Earth Planet. Sci. Lett.* **299**, 466–473 (2010).
38. Kawamura, K. et al. Kinetic fractionation of gases by deep air convection in polar firn. *Atmos. Chem. Phys.* **13**, 11141–11155 (2013).
39. Bereiter, B. et al. Mean global ocean temperatures during the last glacial transition. *Nature* **553**, 39–44 (2018).
40. Severinghaus, J. P. et al. Oxygen-18 of  $\text{O}_2$  records the impact of abrupt climate change on the terrestrial biosphere. *Science* **324**, 1431–1434 (2009).
41. Mitchell, L. et al. Constraints on the Late Holocene anthropogenic contribution to the atmospheric methane budget. *Science* **342**, 964–966 (2013).
42. Ahn, J. H. et al. A high-precision method for measurement of paleoatmospheric  $\text{CO}_2$  in small polar ice samples. *J. Glaciol.* **55**, 499–506 (2009).
43. Bauska, T. K. et al. High-precision dual-inlet IRMS measurements of the stable isotopes of  $\text{CO}_2$  and the  $\text{N}_2\text{O}/\text{CO}_2$  ratio from polar ice core samples. *Atmos. Meas. Tech.* **7**, 3825–3837 (2014).
44. Elderfield, H. et al. Evolution of ocean temperature and ice volume through the mid-Pleistocene climate transition. *Science* **337**, 704–709 (2012).
45. Rohling, E. J. et al. Sea-level and deep-sea-temperature variability over the past 5.3 million years. *Nature* **508**, 477–482 (2014).
46. Lambert, F. et al. Dust-climate couplings over the past 800,000 years from the EPICA Dome C ice core. *Nature* **452**, 616–619 (2008).
47. Siegert, M. J. Glacial–interglacial variations in central East Antarctic ice accumulation rates. *Quat. Sci. Rev.* **22**, 741–750 (2003).
48. Whillans, I. M. & Groote, P. M. Isotopic diffusion in cold snow and firn. *J. Geophys. Res. Atmos.* **90**, 3910–3918 (1985).
49. Etheridge, D. M. et al. Natural and anthropogenic changes in atmospheric  $\text{CO}_2$  over the last 1000 years from air in Antarctic ice and firn. *J. Geophys. Res. Atmos.* **101**, 4115–4128 (1996).
50. Ikeda-Fukazawa, T. et al. Effects of molecular diffusion on trapped gas composition in polar ice cores. *Earth Planet. Sci. Lett.* **229**, 183–192 (2005).
51. Ikeda-Fukazawa, T. et al. Molecular dynamics studies of molecular diffusion in ice Ih. *J. Chem. Phys.* **117**, 3886–3896 (2002).
52. Salamat, A. N. et al. Kinetics of air-hydrate nucleation in polar ice sheets. *J. Cryst. Growth* **223**, 285–305 (2001).
53. Ahn, J. et al.  $\text{CO}_2$  diffusion in polar ice: observations from naturally formed  $\text{CO}_2$  spikes in the Siple Dome (Antarctica) ice core. *J. Glaciol.* **54**, 685–695 (2008).
54. Rempel, A. W. & Wettlaufer, J. S. Isotopic diffusion in polycrystalline ice. *J. Glaciol.* **49**, 397–406 (2003).
55. Vimeux, F. et al. A 420,000 year deuterium excess record from East Antarctica: information on past changes in the origin of precipitation at Vostok. *J. Geophys. Res. Atmos.* **106**, 31863–31873 (2001).
56. Montross, S. et al. Debris-rich basal ice as a microbial habitat, Taylor Glacier, Antarctica. *Geomicrobiol. J.* **31**, 76–81 (2014).
57. Souchez, R. et al. Flow-induced mixing in the GRIP basal ice deduced from the  $\text{CO}_2$  and  $\text{CH}_4$  records. *Geophys. Res. Lett.* **22**, 41–44 (1995).
58. Sowers, T. et al. Elemental and isotopic composition of occluded  $\text{O}_2$  and  $\text{N}_2$  in polar ice. *J. Geophys. Res. Atmos.* **94**, 5137–5150 (1989).
59. Eggelston, S. et al. Evolution of the stable carbon isotope composition of atmospheric  $\text{CO}_2$  over the last glacial cycle. *Paleoceanography* **31**, 434–452 (2016).
60. Veres, D. et al. The Antarctic ice core chronology (AICC2012): an optimized multi-parameter and multi-site dating approach for the last 120 thousand years. *Clim. Past* **9**, 1733–1748 (2013).
61. Bazin, L. et al. An optimized multi-proxy, multi-site Antarctic ice and gas orbital chronology (AICC2012): 120–800 ka. *Clim. Past* **9**, 1715–1731 (2013).

**Acknowledgements** We acknowledge US Ice Drilling Design and Operations (IDDO), driller M. Waszkiewicz, and Ken Borek Air for assistance with the field work. M. Kalk assisted with the  $\text{CO}_2$  measurements. We thank A. Menking and A. Buffen for helping with  $\delta^{13}\text{C}-\text{CO}_2$  measurements. This work received funding from National Science Foundation Grants ANT-1443306 (University of Maine), ANT-1443276 (Oregon State University), NSF-0538630 and ANT-0944343 (Scripps Institution of Oceanography), and ANT-1443263 (Princeton University). Y.Y. acknowledges the Princeton Environmental Institute at Princeton University through the Walbridge Fund, which supported the work upon which this material is partly based.

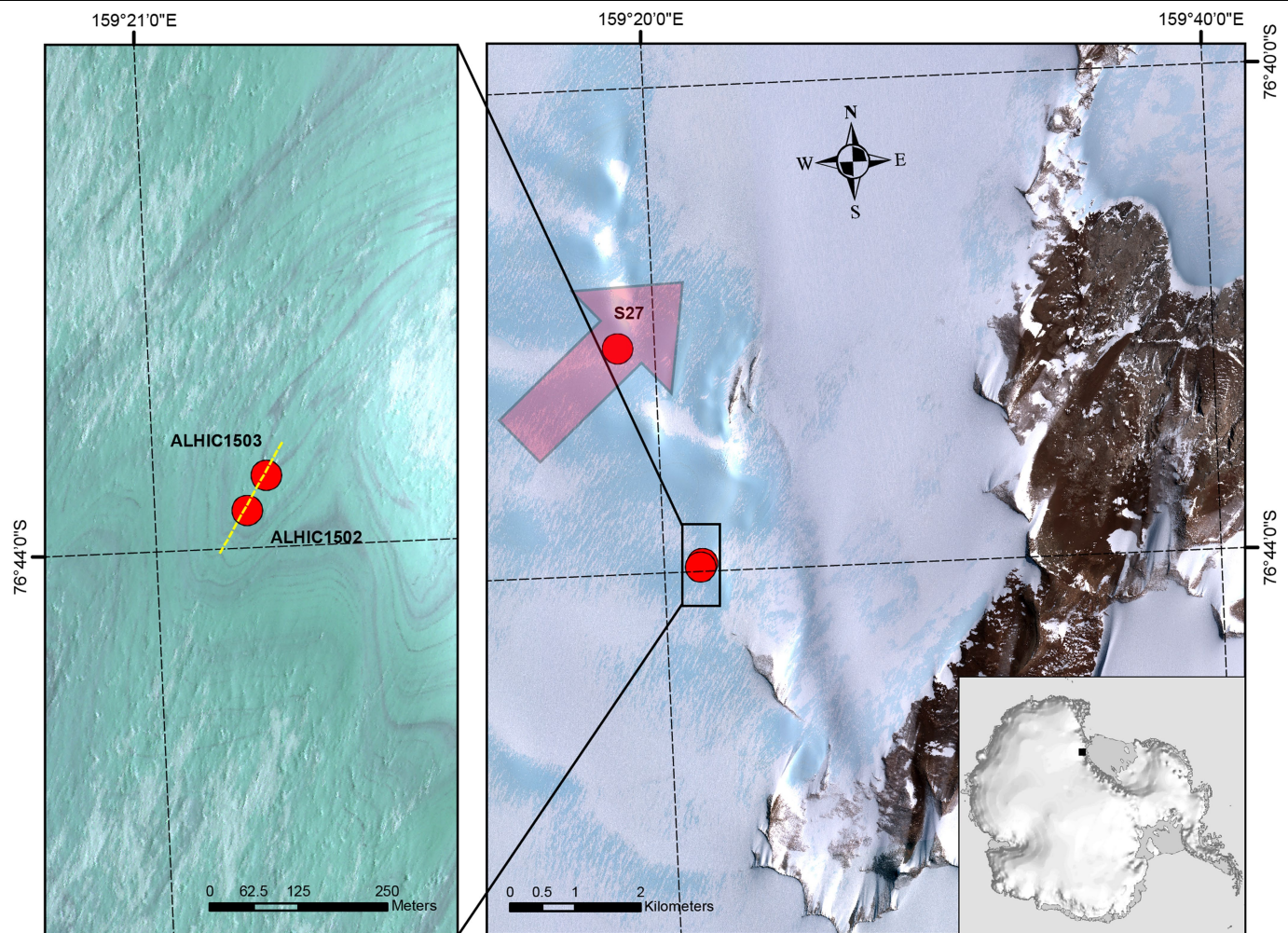
**Author contributions** M.L.B., J.A.H., P.A.M., A.V.K., E.J.B. and J.P.S. designed the research. J.A.H., Y.Y., P.C.K. and S.M. collected the ice core samples. Y.Y. and J.N. performed the  $^{40}\text{Ar}_{\text{atm}}$  and Xe/Kr experiments. Y.Y. analysed the  $\text{O}_2/\text{N}_2/\text{Ar}$  compositions. H.M.C. measured the stable water isotopes. S.M. collected and interpreted the GPR data. Y.Y., J.A.H. and M.L.B. wrote the paper. All authors contributed to the revision of the manuscript before submission.

**Competing interests** The authors declare no competing interests.

### Additional information

**Correspondence and requests for materials** should be addressed to Y.Y.

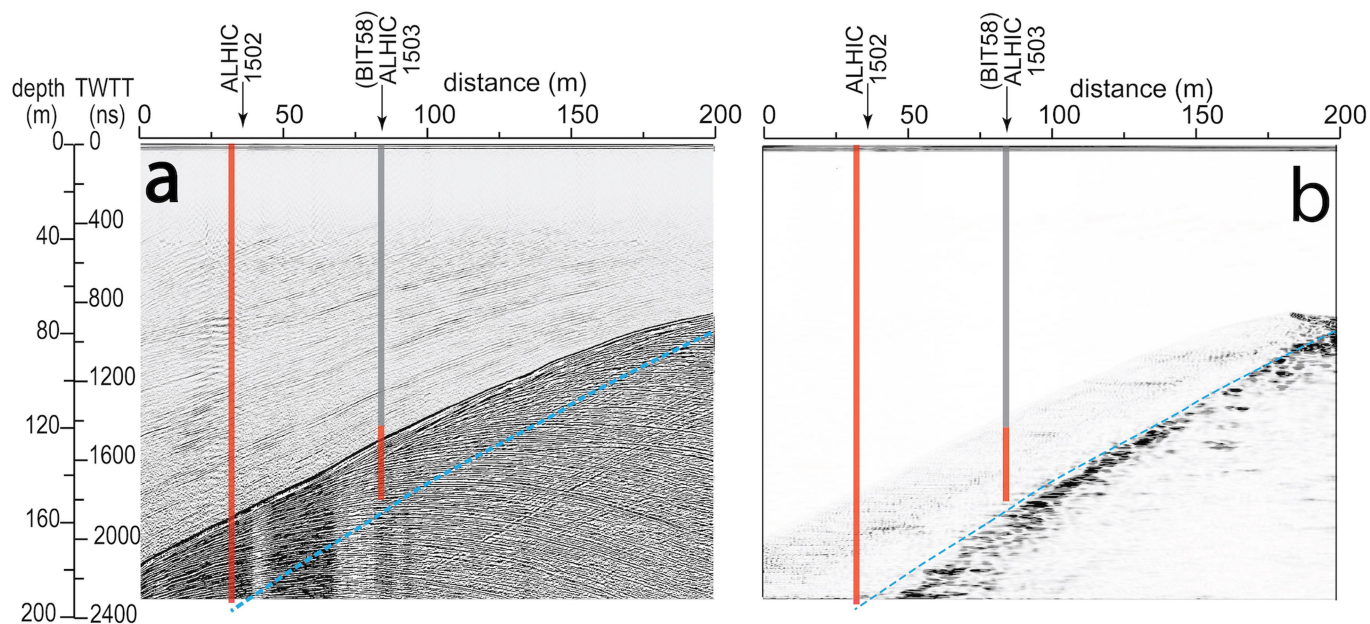
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Satellite imagery of the Allan Hills study area.** Right, WorldView03 colour pan-sharpened imagery (copyright 2011, DigitalGlobe, Inc.) of the main ice field, Allan Hills Blue Ice Area (in true colour mode). Inset, Antarctica with hill shading (source: Australian Antarctic Data Centre, Map 13469; licensed under a Creative Commons Attribution 3.0 Unported License) on which our study area is marked by the black square. The black outcrop in the main image is the Allan Hills nunatak. The red arrow marks the local iceflow.

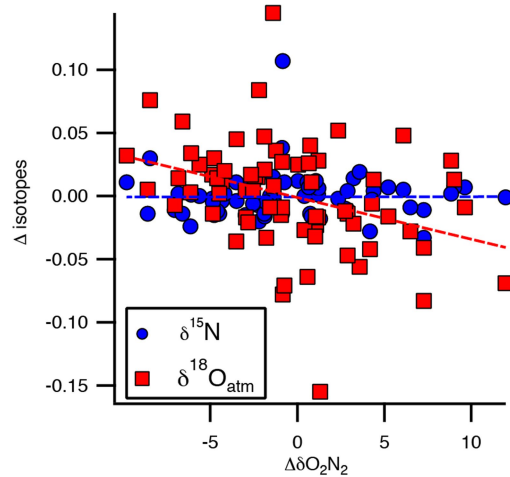
Left, a magnified image of the drilling site from same source file. The imagery is enhanced (gamma-adjusted on ArcGIS 10) to highlight the colour contrast within the blue ice, which now has a greenish hue owing to the colour rendition. The brown lineations in the ice are exposed dust bands, providing a first-order tracer of surface ice stratigraphy. The locations of the cores reported in this work (see text) are marked with red circles. The location of the GPR profile in Extended Data Fig. 2 is shown as a yellow dashed line.



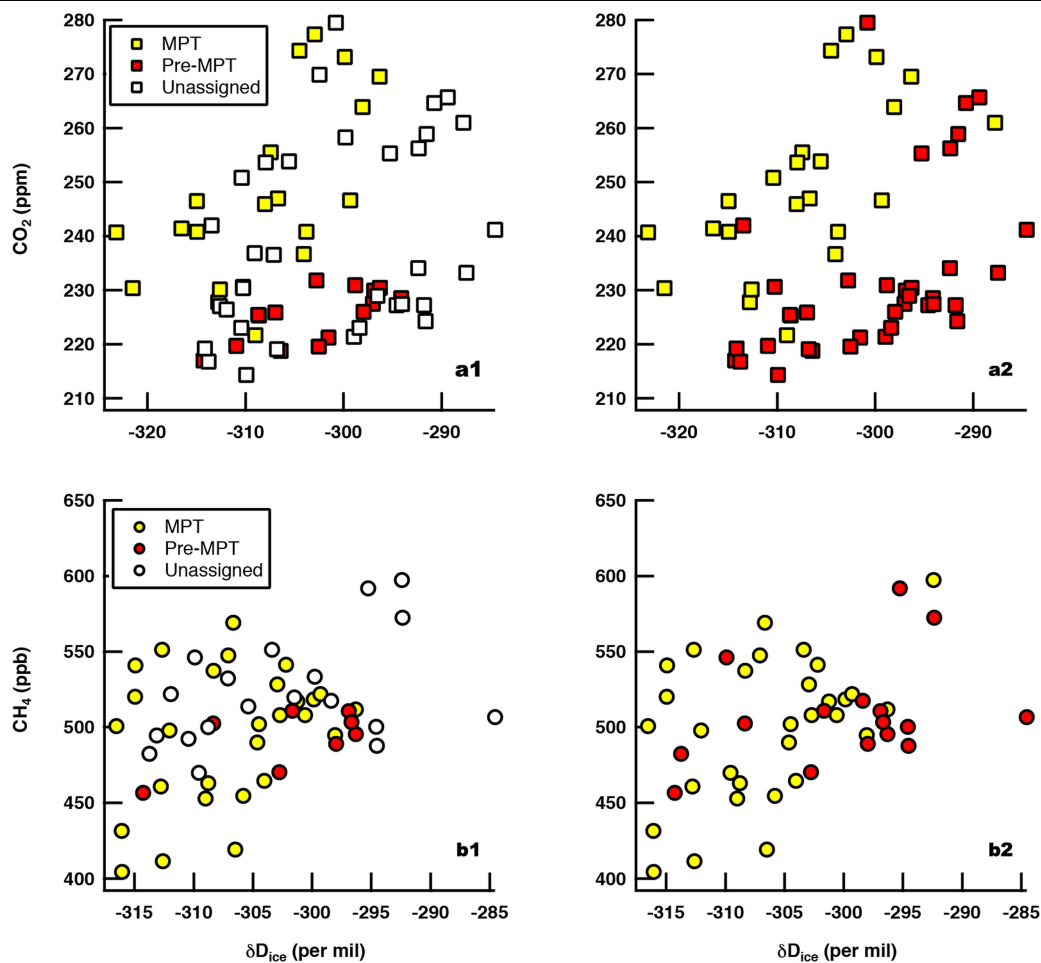


**Extended Data Fig. 2 | GPR profile proceeding from the SW to the NE, crossing (within less than 5 m) ALHIC1502 and ALHIC1503.** The exact transect location is shown in Extended Data Fig. 1. The location and depths of boreholes ALHIC1502 and ALHIC1503 (red bars) and ice drilled previously as BIT58 (grey bar) are indicated. The profile was collected with an 80-MHz MLF antenna in a step-and-collect survey style with a step size of 25 cm and a stacking rate of 64 scans. Standard post-processing steps were applied, including time-zero position correction, distance normalization, 50/110-MHz finite impulse response filter, background removal, and gain adjustment. Depth estimates from two-way travel time (TWTT) and migration use a radar travel velocity of

$0.165 \text{ m ns}^{-1}$ . **a**, Non-migrated radargram showing dipping bed-parallel englacial stratigraphy and a strong dipping apparent bedrock reflection. Dashed blue line indicates the modelled location of the actual bedrock location as shown in **b**. **b**, Migrated radargram showing the 'true' location of the bedrock. Migration quality is poor at this location owing to the steeply dipping slope of the bedrock rise. Comparison between **a** and **b** suggests that the correct bedrock reflector location is translated downward and more steeply dipping than indicated in the non-migrated data. ALHIC1502 and ALHIC1503 borehole depths independently verify this interpretation.

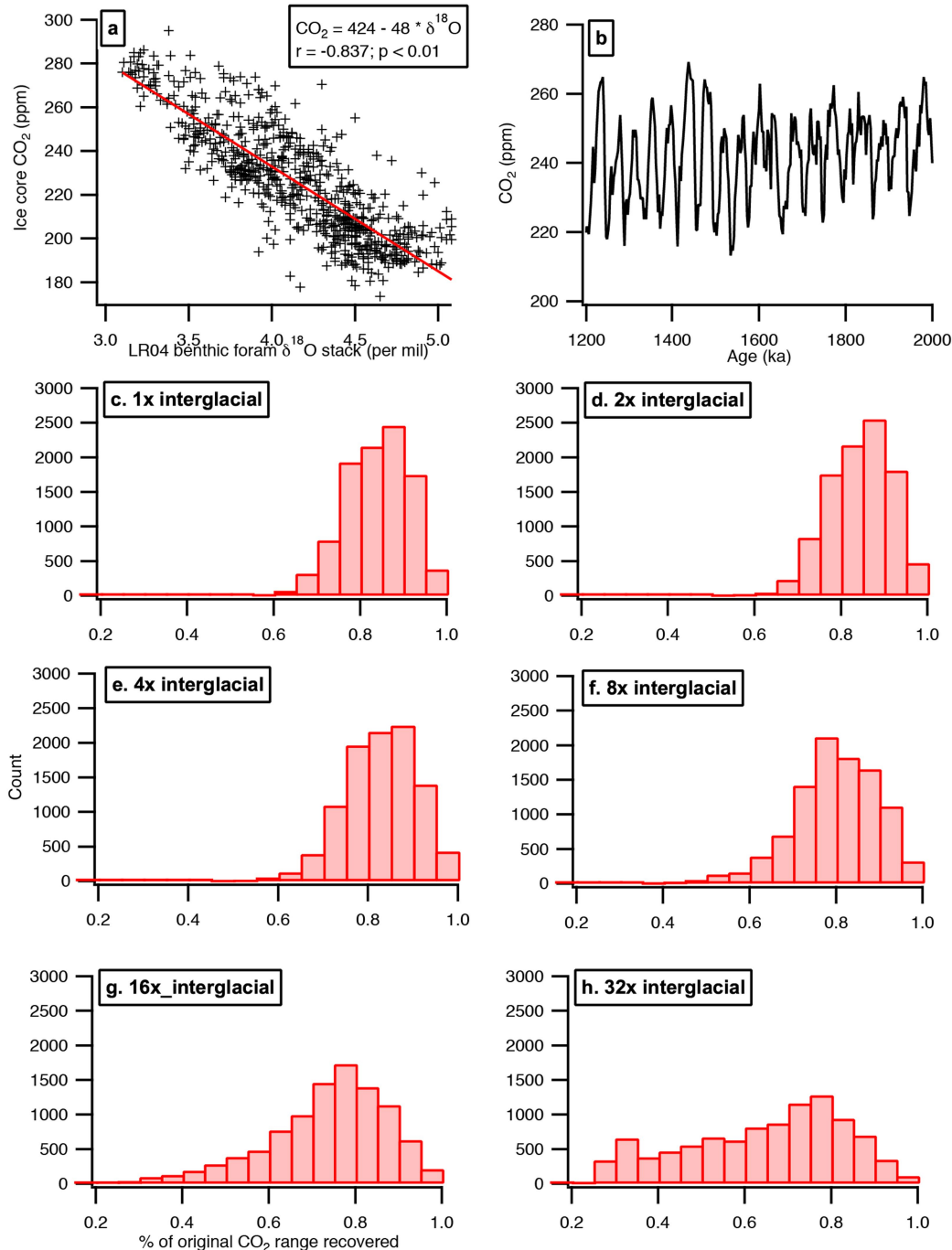


**Extended Data Fig. 3 | Minimal impact of gas loss on oxygen and nitrogen isotopes.** Pair differences of  $\delta^{18}\text{O}_{\text{atm}}$  ( $\Delta\delta^{18}\text{O}_{\text{atm}}$ ; red) and  $\delta^{15}\text{N}$  ( $\Delta\delta^{15}\text{N}$ ; blue) are plotted against  $\Delta\delta\text{O}_2/\text{N}_2$  in ALHIC1502 and ALHIC1503 samples. The difference between two replicate samples is attributable to gas loss. The very weak dependence of  $\Delta\delta^{18}\text{O}_{\text{atm}}$  and  $\Delta\delta^{15}\text{N}$  on  $\Delta\delta\text{O}_2/\text{N}_2$  suggests that gas isotopes are relatively insensitive to gas loss fractionation in Allan Hills ice.



**Extended Data Fig. 4 | Age assignment of CO<sub>2</sub> and CH<sub>4</sub> samples.** Using the two age models discussed in the text—the conservative approach (top and bottom left) and the proximity approach (top and bottom right)—CO<sub>2</sub> and CH<sub>4</sub> are

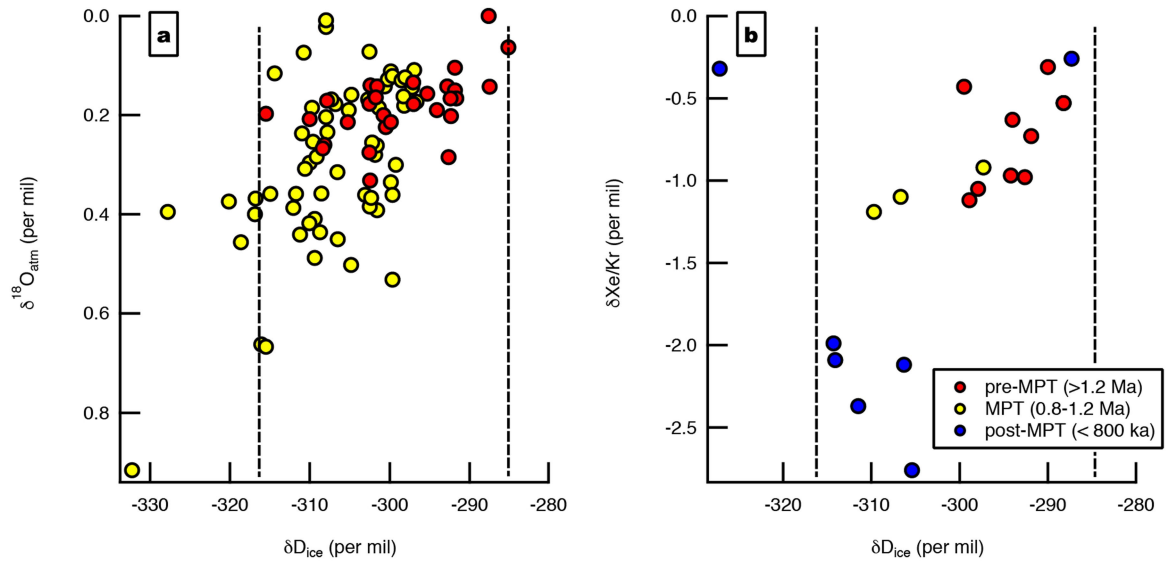
plotted against δD<sub>ice</sub>. Both of the age models show reduced range in the MPT and pre-MPT ice. We note, however, that the highest CO<sub>2</sub> and CH<sub>4</sub> values fall into the unassigned category in the strictest age model.



**Extended Data Fig. 5 | Evaluating the fraction of CO<sub>2</sub> range captured by 37 discrete samples.** **a**, Ice core CO<sub>2</sub> record between 0 and 800 ka versus LR04 benthic foram δ<sup>18</sup>O stack synchronized onto the AICC2012 timescale<sup>60,61</sup>. The result of a linear regression is shown and used to calculate the synthetic CO<sub>2</sub> record between 1.2 and 2.0 Ma. **b**, A synthetic CO<sub>2</sub> record between 1.2 and 2.0 Ma reconstructed from LR04 δ<sup>18</sup>O and the regression parameters calculated in **a**. The range of CO<sub>2</sub> in this synthetic time series is 213–269 ppm. **c–h**, Fraction of the recovered CO<sub>2</sub> variability (observed range/true range) of the synthetic CO<sub>2</sub>

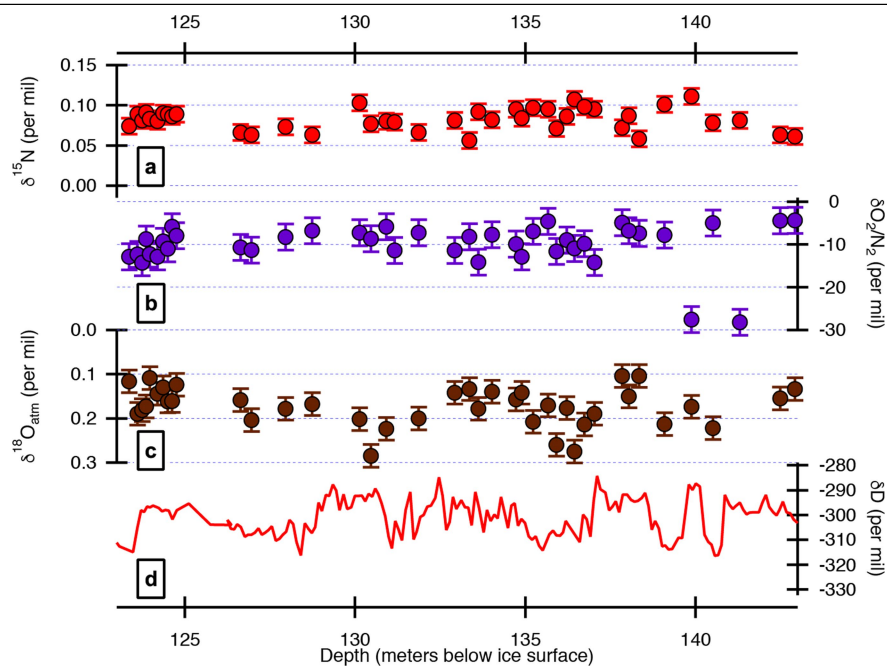
time series (see Extended Data Fig. 5b) by 37 randomly selected samples, given different preferential preservation of interglacial ice. Here the term ‘interglacial’ is operationally defined as any sample with greater than 250 ppm CO<sub>2</sub>. Different multipliers indicate the multiple occurrence of the ‘interglacial’ ice to simulate varying degrees of ice preservation biases. This analysis shows that when the presence of interglacial ice is no more than eight times greater than that of glacial ice (**c–f**), we could expect 37 random samples to be likely to capture 66–98% (95% confidence interval) of the CO<sub>2</sub> range in the ice.





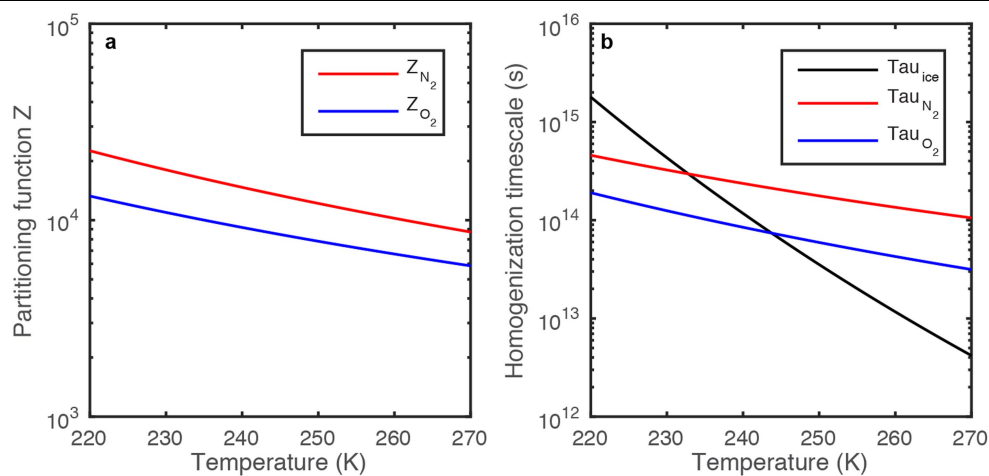
**Extended Data Fig. 6 | Evaluating the fraction of  $\delta^{18}\text{O}_{\text{atm}}$  and  $\delta\text{Xe/Kr}$  range captured by discrete samples. a, b,** Allan Hills  $\delta^{18}\text{O}_{\text{atm}}$  (a) and  $\delta\text{Xe/Kr}$  (b) are plotted against the  $\delta\text{D}_{\text{ice}}$  of the same depth, colour coded according to their age units. Vertical dashed lines represent the range of  $\delta\text{D}_{\text{ice}}$  observed in the 40k-world ice (–284 to –316‰). Notably, the range of  $\delta\text{D}_{\text{ice}}$  associated with nine  $\delta\text{Xe/Kr}$  values from the 40k world is only about 40% of the entire  $\delta\text{D}_{\text{ice}}$  range in our 40k ice samples (–288 to –300‰); low  $\delta\text{D}_{\text{ice}}$  values (less than –300‰) are

missing from the  $\delta\text{Xe/Kr}$  samples. Thus, we do not expect the nine Xe/Kr samples to fully capture the range of Xe/Kr ratios in the 40k-world ice. On the other hand, the co-depth  $\delta\text{D}_{\text{ice}}$  values of  $\delta^{18}\text{O}_{\text{atm}}$  samples occupy 97% of the total range, implying that 29  $\delta^{18}\text{O}_{\text{atm}}$  samples are likely to cover 70% of the variability preserved in the 40k-world ice, barring any diffusive smoothing of the  $\delta^{18}\text{O}_{\text{atm}}$  records.



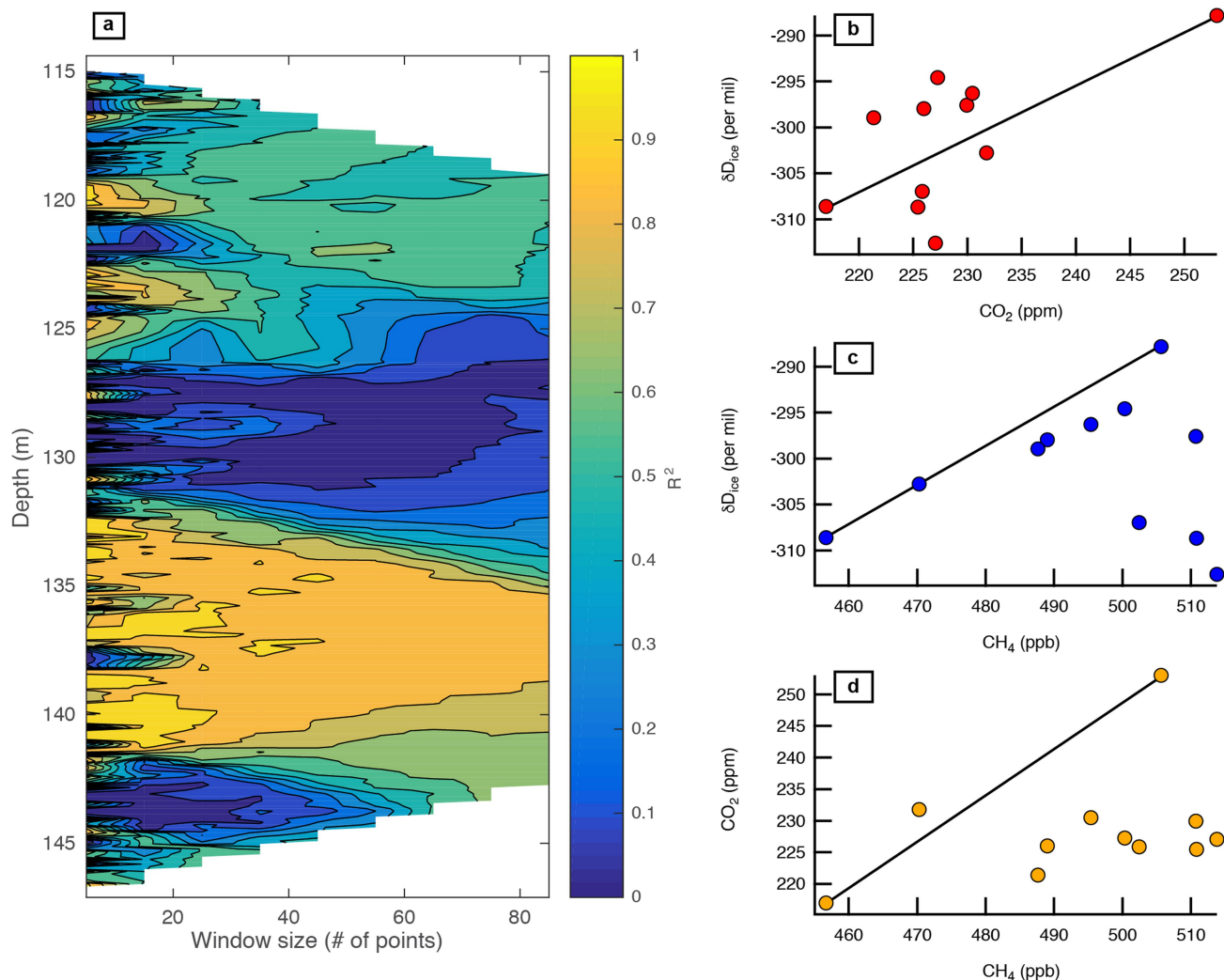
**Extended Data Fig. 7 | Gas and ice properties in the interval between 123 and 143 m in ALHIC1503. a,  $\delta^{15}\text{N}$ ; b,  $\delta\text{O}_2/\text{N}_2$ ; c,  $\delta^{18}\text{O}_{\text{atm}}$ ; d,  $\delta\text{D}_{\text{ice}}$ .** The error bars associated with gas properties represent the pooled standard deviation ( $1\sigma$ ) of

all measurements: 0.010‰ for  $\delta^{15}\text{N}$ , 3.063‰ for  $\delta\text{O}_2/\text{N}_2$ , and 0.026‰ for  $\delta^{18}\text{O}_{\text{atm}}$ . Note that the range of  $\delta^{18}\text{O}_{\text{atm}}$  is less than 0.2‰. By comparison, glacial-to-interglacial  $\delta^{18}\text{O}_{\text{atm}}$  variability in the 100-kyr climate cycles is about 1.5‰<sup>21</sup>.



**Extended Data Fig. 8 | Evaluating the effect of diffusive mixing on ice and gas properties.** **a**, Partitioning function  $Z$  (ratio of the number gas molecules in the gas phase to that in the ice phase) of  $O_2$  (blue) and  $N_2$  (red) as a function of temperature. **b**, Characteristic time scale of the self-diffusion of water molecules

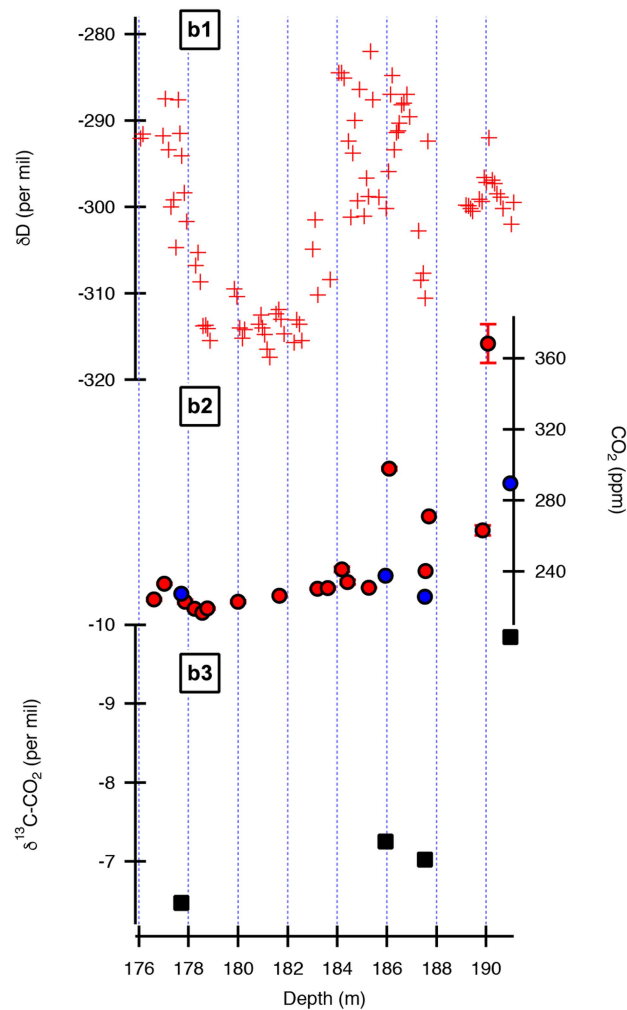
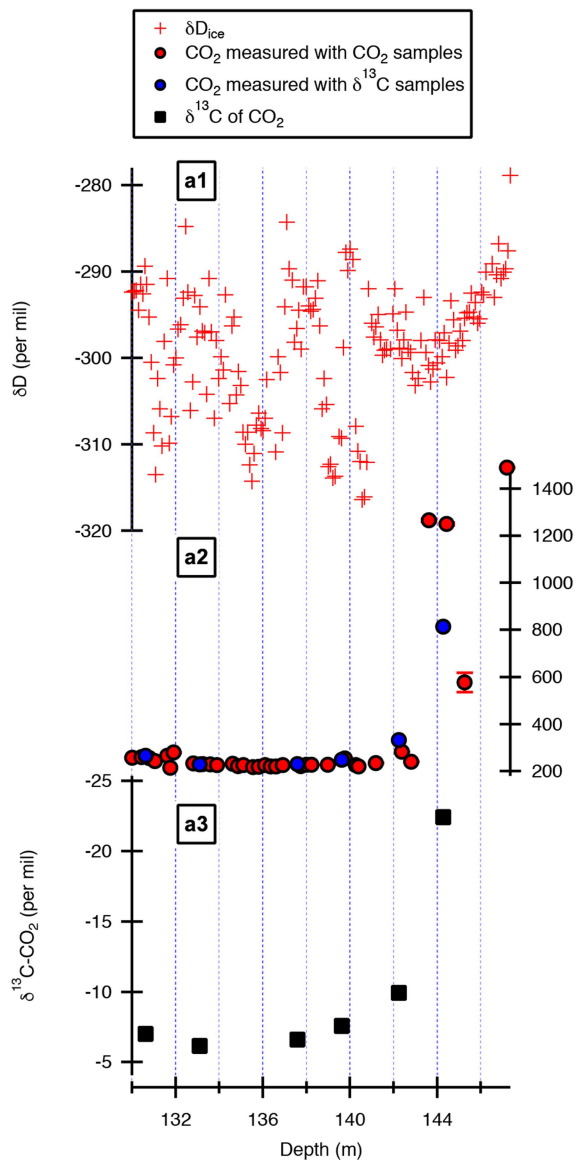
in ice (black), of  $O_2$  permeation in ice (blue), and of  $N_2$  permeation in ice (red), plotted as a function of temperature. Note that for water molecules the length scale of diffusion ( $L$ ) is 0.1 m, whereas for  $N_2$  and  $O_2$   $L = 0.5$  m, reflective of their respective sampling resolutions.



**Extended Data Fig. 9 | Evaluating the possibility of flow-induced mixing in Allan Hills ice cores.** **a**, A contour map shows the correlation coefficient ( $R^2$ ) between  $\delta D_{ice}$  and  $d$  in core ALHIC1503 as a function of the number of adjacent stable water isotope data points included in the calculations, from 5 to 85 with a step of 10. The sampling resolution is approximately 10 cm. The high correlation coefficient ( $>0.7$ ) between 132 and 142 m raises the suspicion that mixing has taken place in this interval. However, the possibility of mixing is not supported

by  $CO_2$  and  $CH_4$ . **b–d**, Cross-plots of  $\delta D_{ice}$ – $CO_2$  (**b**),  $\delta D_{ice}$ – $CH_4$  (**c**), and  $CO_2$ – $CH_4$  (**d**) in intervals between 132 and 140 m in ALHIC1503. The black solid lines are mixing lines based on two end members at 135.30 m and at 139.76 m, where the minimum and maximum  $CO_2$  values are measured, respectively. Most of the points do not fall onto the mixing line, implying that two-end-member mixing alone cannot explain the high correlation between  $\delta D_{ice}$  and  $d$ .





**Extended Data Fig. 10 | Respiration in basal ice revealed by  $\delta^{13}C$  of  $CO_2$  in ALHIC1503 and ALHIC1502.** **a**, ALHIC1503; **b**, ALHIC1502.  $\delta D_{ice}$  (top),  $CO_2$  (middle), and  $\delta^{13}C$  of  $CO_2$  (bottom) in Allan Hills ice cores are plotted against depth.  $\delta D_{ice}$  and  $CO_2$  (both measured independently in smaller samples,

plotted in red, and along with  $\delta^{13}C$ , plotted in blue) are also shown for comparison. Note the marked scale difference for  $CO_2$  and  $\delta^{13}C$  between ALHIC1503 and ALHIC1502.

# Widespread global increase in intense lake phytoplankton blooms since the 1980s

<https://doi.org/10.1038/s41586-019-1648-7>

Jeff C. Ho<sup>1,2\*</sup>, Anna M. Michalak<sup>1\*</sup> & Nima Pahlevan<sup>3,4</sup>

Received: 11 January 2018

Accepted: 6 August 2019

Published online: 14 October 2019

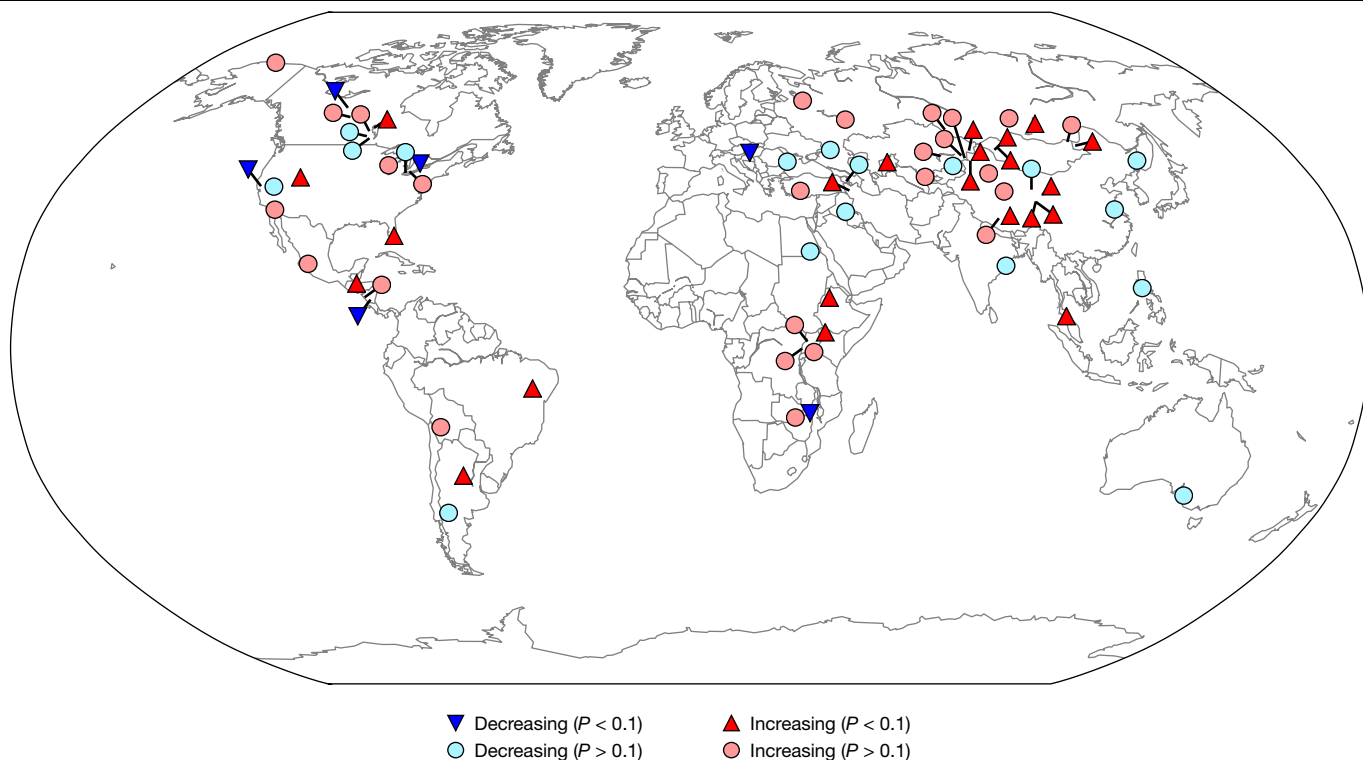
Freshwater blooms of phytoplankton affect public health and ecosystem services globally<sup>1,2</sup>. Harmful effects of such blooms occur when the intensity of a bloom is too high, or when toxin-producing phytoplankton species are present. Freshwater blooms result in economic losses of more than US\$4 billion annually in the United States alone, primarily from harm to aquatic food production, recreation and tourism, and drinking-water supplies<sup>3</sup>. Studies that document bloom conditions in lakes have either focused only on individual or regional subsets of lakes<sup>4–6</sup>, or have been limited by a lack of long-term observations<sup>7–9</sup>. Here we use three decades of high-resolution Landsat 5 satellite imagery to investigate long-term trends in intense summertime near-surface phytoplankton blooms for 71 large lakes globally. We find that peak summertime bloom intensity has increased in most (68 per cent) of the lakes studied, revealing a global exacerbation of bloom conditions. Lakes that have experienced a significant ( $P < 0.1$ ) decrease in bloom intensity are rare (8 per cent). The reason behind the increase in phytoplankton bloom intensity remains unclear, however, as temporal trends do not track consistently with temperature, precipitation, fertilizer-use trends or other previously hypothesized drivers. We do find, however, that lakes with a decrease in bloom intensity warmed less compared to other lakes, suggesting that lake warming may already be counteracting management efforts to ameliorate eutrophication<sup>10,11</sup>. Our findings support calls for water quality management efforts to better account for the interactions between climate change and local hydrological conditions<sup>12,13</sup>.

The reported incidence of toxic phytoplankton blooms has risen considerably over the past half-century<sup>14</sup>. While it is generally understood that nutrient loading drives phytoplankton blooms<sup>15</sup>, the degree to which bloom conditions are changing globally and the factors that drive these changes among multiple interacting stressors<sup>16</sup> are still uncertain<sup>17</sup>. An understanding of global patterns, trends and drivers is necessary, however, for designing effective management and remediation strategies<sup>18</sup>. Whereas past studies synthesizing information on the long-term trends in phytoplankton blooms of lakes have been limited by data availability, recent advances in cloud-based parallel computing have made it possible to leverage high-resolution freely accessible satellite imagery over large areas, enabling the study of long-term environmental trends on a global scale<sup>19,20</sup>.

Here, we take advantage of these advances to generate a long-term record of intense, near-surface phytoplankton blooms for dozens of large lakes across the globe. We use data from the Landsat 5 satellite to generate time series of peak summer bloom intensity from 1984 to 2012 for 71 lakes in 33 countries across 6 continents (Fig. 1). In total, the data span 30,922 scenes and 72.6 billion lake pixels. The study lakes span a broad range of physical characteristics and degree of anthropogenic impacts (Supplementary Table 1; see Methods for a full description of the implemented approach). Seasonal peak bloom intensity for a given

lake and year is defined based on the maximum observed lake-wide near-infrared signal magnitude, with a first-order correction of atmospheric interference using the shortwave-infrared signal<sup>21</sup>. Remotely sensed observations within the near-infrared part of the electromagnetic spectrum are sensitive to intense, near-surface algal blooms (see Methods). An initial superset of 154 lakes was selected based on their inclusion in previous studies that leveraged remote sensing by satellites<sup>22,23</sup>, thus reducing the likelihood that persistent cloudiness obscured the images. These lakes all have surface areas of more than 100 km<sup>2</sup>; globally, lakes within this size range contain approximately 95% of all water stored in lakes<sup>24</sup>. Data of lakes for which little signal was observed throughout the study period, as well as data of lakes for which the signal was far outside the range over which the original algorithm was designed<sup>21</sup>, were removed. A smaller number of additional lakes were removed due to previously documented evidence of a lack of phytoplankton blooms. Of the final selected lakes, 38 have a documented presence of harmful cyanobacterial species, while the rest show evidence of other phytoplankton species (10 lakes) or no reported evidence of blooms (23 lakes). Given the heterogeneity in lake characteristics, the time series of the interannual bloom intensity for each lake is normalized by its own long-term mean and s.d. to assess the relative change in bloom intensity over time. This approach eliminates the need to compare absolute

<sup>1</sup>Department of Global Ecology, Carnegie Institution for Science, Stanford, CA, USA. <sup>2</sup>Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA. <sup>3</sup>NASA Goddard Space Flight Center, Greenbelt, MD, USA. <sup>4</sup>Science Systems and Applications Inc, Lanham, MD, USA. \*e-mail: jeffho@stanford.edu; michalak@stanford.edu



**Fig. 1 | Global distribution of lake bloom intensity trends shows that the peak summertime bloom intensity has increased since the 1980s.** The map shows bloom intensity trends for all 71 study lakes for the period 1984–2012 (Supplementary Table 1). Colours and symbols indicate whether the bloom

intensity decreased (blue) or increased (red), and whether the trend is statistically significant (triangles for  $P < 0.1$ ; circles for  $P > 0.1$ ). The base map was generated using Generic Mapping Tools<sup>33</sup>.

magnitudes across lakes, which has been an important barrier to past syntheses across lakes<sup>25</sup>.

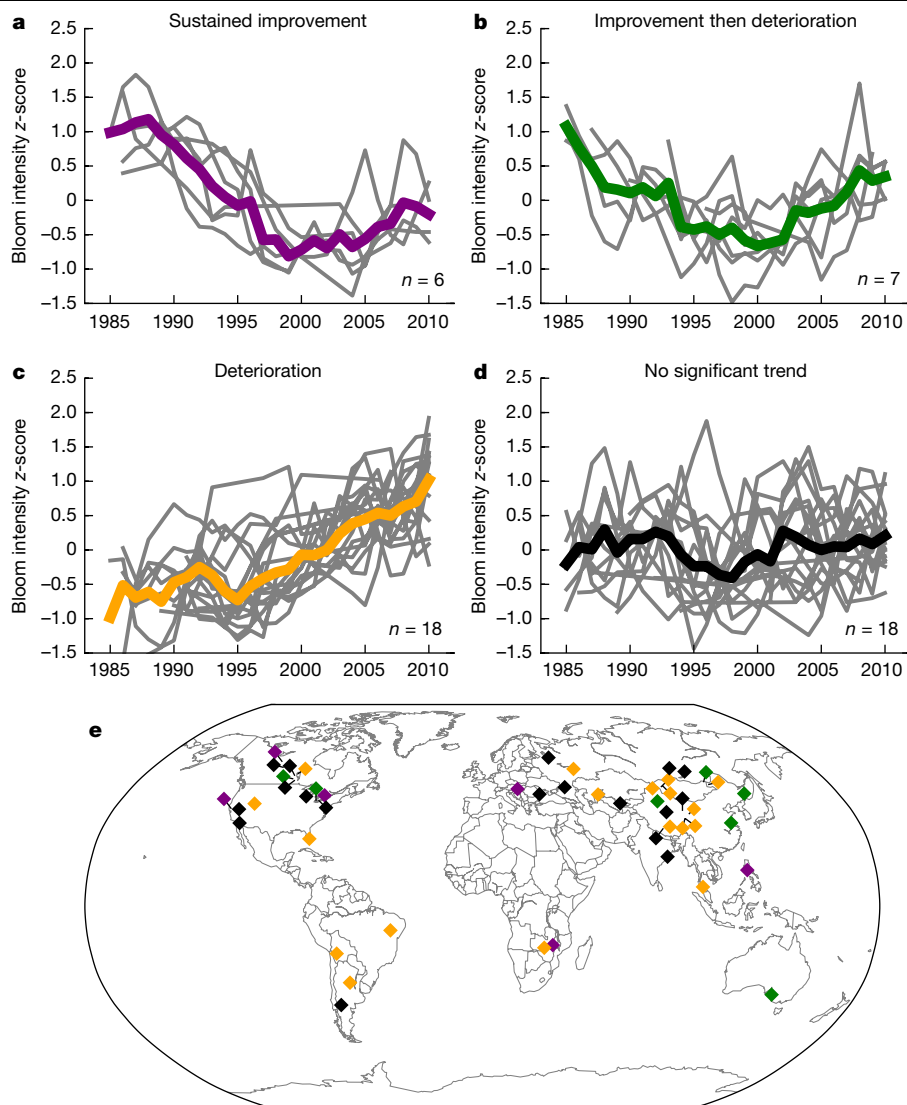
We find that the implemented algorithm is able to successfully capture previously documented spatial gradients in the severity of phytoplankton blooms within individual lakes and temporal trends in phytoplankton bloom intensity for specific lakes (Extended Data Fig. 1 and Methods). Using simulations of atmospheric radiative transfer, we also find that the algorithm is insensitive to reported variations in Landsat 5 orbit or image radiometric quality, primarily owing to the strong signal that arises from the intense, near-surface blooms identified in study lakes (see Supplementary Information). These results suggest that a single algorithm can indeed identify intense phytoplankton blooms despite the large differences in optical properties across lakes<sup>26</sup>, as long as the focus is on interannual rather than inter-lake variability. This lends support to the approach implemented in this study for tracking long-term trends globally. We then used all 71 study lakes to assess global trends in summertime peak phytoplankton bloom intensity. We also used a subset of 49 lakes with at least 14 years of data to explore more detailed historical temporal patterns in phytoplankton bloom conditions, for which the 14-year threshold was selected on the basis of previously published studies on global lake temperatures<sup>22,27</sup>.

We find that peak summertime phytoplankton bloom intensity has increased in more than two-thirds of study lakes since the 1980s (48 out of 71 lakes) (Fig. 1). Increases in bloom intensity are statistically significant for close to a third of all lakes ( $P < 0.1$  for 22 out of 71 lakes), whereas only 6 lakes exhibited a statistically significant decrease in intensity ( $P < 0.1$ ). A similar proportion of lakes has an increasing bloom intensity among those with a documented presence of cyanobacteria (24 out of 38 lakes) compared to lakes without cyanobacteria (24 out of 33 lakes), and the proportion of lakes with increases in bloom intensity is also consistent across lakes with different areas, volumes, mean and maximum depths, and latitudes (see Supplementary Table 1 and Supplementary

Information). These results suggest that the observed trends are widespread globally and across lake types, in contrast to previous hypotheses of differential impacts as a function of latitude<sup>28</sup> or morphometry<sup>29</sup>. This finding provides a global perspective that is consistent with surveys of sedimentary records across temperate–subarctic lakes<sup>6</sup> that show sharp increases in the concentrations of cyanobacterial pigments after 1985. This finding also corroborates putative trends of increasing harmful cyanobacterial blooms globally<sup>17</sup>, and counters the hypothesis that increased reporting of toxic blooms is instead a by-product of increased scientific attention<sup>30</sup>.

We find that lake phytoplankton bloom histories follow one of four prototypical pathways, termed here ‘sustained improvement’, ‘improvement then deterioration’, ‘deterioration’ and ‘no significant trend’ (Fig. 2a–d and Methods). The two pathways that include deteriorating conditions reveal that increases in peak bloom intensity occurred predominantly in the latter half of the study period (Fig. 2b, c). For example, three-quarters of study lakes (51 out of 68) with sufficient data for the second half of the study period (1998–2012) exhibited an increase in bloom intensity during this period, whereas only a third (22 out of 66) experienced an increase during the first half (1984–1997). The reason behind the temporal coherence of changes in phytoplankton bloom intensity remains unclear, as temporal trends do not track consistently with temperature, precipitation, fertilizer-use trends, satellite data availability or geomorphological characteristics of individual lakes (Extended Data Figs. 2–5 and Supplementary Information), nor are there widespread trends in the seasonal timing of peak bloom intensity (see Supplementary Information).

We find that although lakes that exhibited sustained improvement were rare ( $n = 6$ ), they experienced less warming (or more cooling) relative to those that exhibited improvement then deterioration ( $P = 0.09$ ; Fig. 3 and Extended Data Fig. 6), suggesting that lake warming may have counteracted management efforts in the latter group. This finding suggests that nutrient reduction targets based on historical relationships



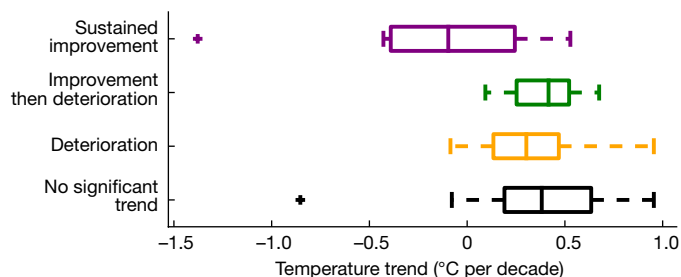
**Fig. 2 | Lake bloom histories follow one of four prototypical pathways.** **a–d**, Time series for lakes with at least 14 years of data ( $n = 49$ ) categorized by historical pathway. Grey lines show 5-year moving averages of normalized bloom intensity, with coloured lines showing pathway averages across lakes. The time

series of the bloom intensity z-score for each lake is calculated using its own historical mean and s.d. **e**, Global distribution of lake pathways. The base map was generated using Generic Mapping Tools<sup>33</sup>.

between bloom severity and nutrient loading may have to be revised in the context of climate change, as has been hypothesized<sup>11</sup>. Generalizing the impact of warming across a wide range of lakes is inadvisable, however, as trends across the full lake ensemble showed little direct

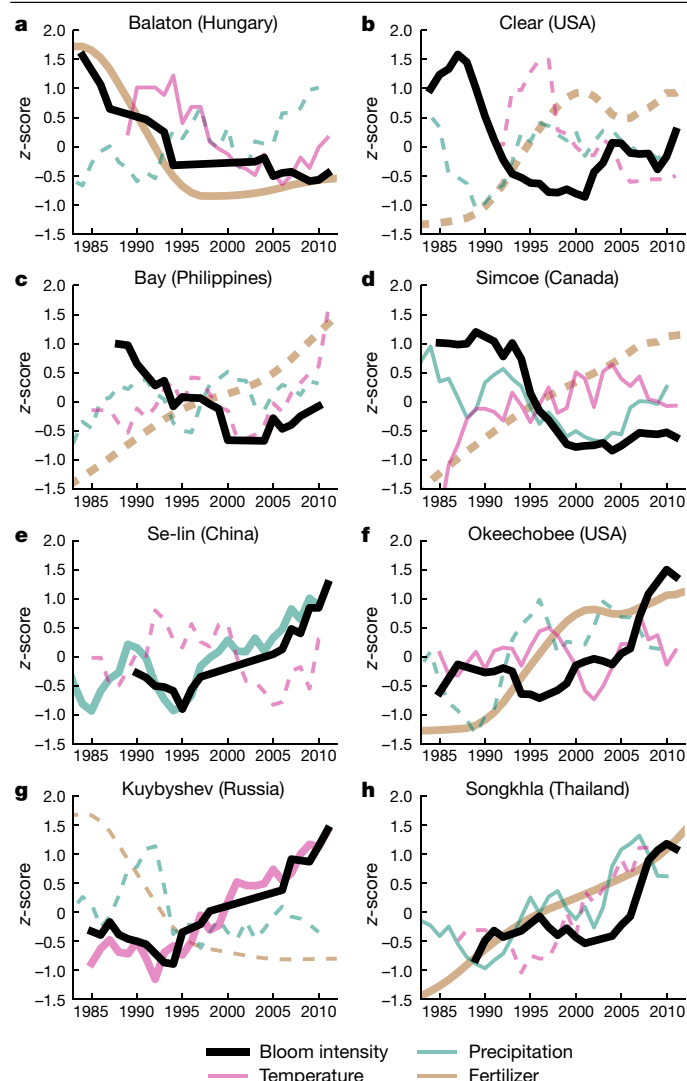
correlation with temperature (Fig. 4, Extended Data Figs. 2, 3 and Supplementary Information). Rather, these findings suggest that the effects of global lake warming differ depending on lake-specific characteristics<sup>31</sup>, and highlight the importance of assessing the role of lake attributes in modulating the impact of temperature on nutrient–phytoplankton relationships<sup>32</sup>.

Overall, this study provides a global view of trends in intense lacustrine near-surface phytoplankton blooms over the past three decades. We examine bloom histories for lakes with widely differing characteristics and geographical locations, and demonstrate the promise of long-term satellite observations for tracking intense bloom conditions across a heterogeneous set of systems to augment geographically and temporally limited in situ monitoring efforts. Our results corroborate the putative reported increase in bloom occurrence and intensity globally, and highlight that lakes that have exhibited a long-term decrease in bloom intensity are rare. Results further show that sustained decreases in bloom intensity are more likely to have occurred in lakes with little or no warming, suggesting that rising lake temperatures may hamper environmental recovery, and illustrating the importance of identifying factors that make some lakes more susceptible to the effects of warming.



**Fig. 3 | Lakes that experienced improvements in bloom conditions tend to have experienced little to no warming.** Box plots of the water temperature trend (1985–2012) binned by lake historical pathway. Each box extends from the first to the third quartile values, with a line at the median. The whiskers extend to  $1.5 \times$  the interquartile range from the edges of the box. The plus symbols show outlier values past the end of the whiskers.





**Fig. 4 | Lake bloom histories show no consistent correspondence with temperature, precipitation and fertilizer use.** **a–h**, Five-year moving averages of normalized near-surface bloom intensity, summer lake temperatures, and total precipitation and fertilizer application rate over the watershed for eight prototypical lakes. **a–d**, Lakes follow the sustained improvement pathway. **e–h**, Lakes follow the deterioration pathway. Thicker temperature, precipitation and fertilizer lines indicate that the Pearson correlation coefficient with bloom intensity is significant ( $P < 0.1$ ). Dashed lines indicate anti-correlations.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1648-7>.

- Pick, F. R. Blooming algae: a Canadian perspective on the rise of toxic cyanobacteria. *Can. J. Fish. Aquat. Sci.* **73**, 1149–1158 (2016).

- Ndilela, L. L., Oberholster, P. J., Van Wyk, J. H. & Cheng, P. H. An overview of cyanobacterial bloom occurrences and research in Africa over the last decade. *Harmful Algae* **60**, 11–26 (2016).
- Kudela, R. M. et al. *Harmful Algal Blooms. A Scientific Summary For Policy Makers* (IOC/UNESCO, 2015).
- Hampton, S. E. et al. Sixty years of environmental change in the world's largest freshwater lake – Lake Baikal, Siberia. *Glob. Change Biol.* **14**, 1947–1958 (2008).
- Duan, H. et al. Two-decade reconstruction of algal blooms in China's Lake Taihu. *Environ. Sci. Technol.* **43**, 3522–3528 (2009).
- Taranu, Z. E. et al. Acceleration of cyanobacterial dominance in north temperate-subarctic lakes during the Anthropocene. *Ecol. Lett.* **18**, 375–384 (2015).
- Carvalho, L. et al. Sustaining recreational quality of European lakes: minimizing the health risks from algal blooms through phosphorus control. *J. Appl. Ecol.* **50**, 315–323 (2013).
- Beaulieu, M., Pick, F. & Gregory-Eaves, I. Nutrients and water temperature are significant predictors of cyanobacterial biomass in a 1147 lakes data set. *Limnol. Oceanogr.* **58**, 1736–1746 (2013).
- Kosten, S. et al. Warmer climates boost cyanobacterial dominance in shallow lakes. *Glob. Change Biol.* **18**, 118–126 (2012).
- Posch, T., Köster, O., Salcher, M. M. & Pernthaler, J. Harmful filamentous cyanobacteria favoured by reduced water turnover with lake warming. *Nat. Clim. Change* **2**, 809–813 (2012).
- Paerl, H. W. et al. Mitigating cyanobacterial harmful algal blooms in aquatic ecosystems impacted by climate change and anthropogenic nutrients. *Harmful Algae* **54**, 213–222 (2016).
- Winder, M. Lake warming mimics fertilization. *Nat. Clim. Change* **2**, 771–772 (2012).
- Paerl, H. W. & Huisman, J. Blooms like it hot. *Science* **320**, 57–58 (2008).
- Carmichael, W. in *Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs* (ed. Hudnell, H. K.) 105–125 (Springer-Verlag, 2008).
- Schindler, D. W., Carpenter, S. R., Chapra, S. C., Hecky, R. E. & Orihel, D. M. Reducing phosphorus to curb lake eutrophication is a success. *Environ. Sci. Technol.* **50**, 8923–8929 (2016).
- Winter, J. G., Young, J. D., Landre, A., Stainsby, E. & Jarjanazi, H. Changes in phytoplankton community composition of Lake Simcoe from 1980 to 2007 and relationships with multiple stressors. *J. Great Lakes Res.* **37**, 63–71 (2011).
- Huisman, J. et al. Cyanobacterial blooms. *Nat. Rev. Microbiol.* **16**, 471–483 (2018).
- McCrackin, M. L., Jones, H. P., Jones, P. C. & Moreno-Mateos, D. Recovery of lakes and coastal marine ecosystems from eutrophication: a global meta-analysis. *Limnol. Oceanogr.* **62**, 507–518 (2017).
- Gorelick, N. et al. Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).
- Zhu, Z. et al. Benefits of the free and open Landsat data policy. *Remote Sens. Environ.* **224**, 382–385 (2019).
- Ho, J. C., Stumpf, R. P., Bridgeman, T. B. & Michalak, A. M. Using Landsat to extend the historical record of lacustrine phytoplankton blooms: a Lake Erie case study. *Remote Sens. Environ.* **191**, 273–285 (2017).
- Schneider, P. & Hook, S. J. Space observations of inland water bodies show rapid surface warming since 1985. *Geophys. Res. Lett.* **37**, L22405 (2010).
- Sharma, S. et al. A global database of lake surface temperatures collected by in situ and satellite methods from 1985–2009. *Sci. Data* **2**, 150008 (2015).
- Messenger, M. L., Lehner, B., Grill, G., Nedeva, I. & Schmitt, O. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nat. Commun.* **7**, 13603 (2016).
- Hampton, S. E. Understanding lakes near and far. *Science* **342**, 815–816 (2013).
- Spyrakos, E. et al. Optical types of inland and coastal waters. *Limnol. Oceanogr.* **63**, 846–870 (2018).
- O'Reilly, C. M. et al. Rapid and highly variable warming of lake surface waters around the globe. *Geophys. Res. Lett.* **42**, 10,773–10,781 (2015).
- Downing, J. in *Global Environmental Change* (ed. Freedman, B.) 221–229 (Springer, 2014).
- Kraemer, B. M. et al. Morphometry and average temperature affect lake stratification responses to climate change. *Geophys. Res. Lett.* **42**, 4,981–4,988 (2015).
- Fristachi, A. et al. in *Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs* (ed. Hudnell, H. K.) 45–103 (Springer-Verlag, 2008).
- Adrian, R. et al. Lakes as sentinels of climate change. *Limnol. Oceanogr.* **54**, 2,283–2,297 (2009).
- Rigosi, A., Carey, C. C., Ibelings, B. W. & Brookes, J. D. The interaction between climate warming and eutrophication to promote cyanobacteria is dependent on trophic state and varies among taxa. *Limnol. Oceanogr.* **59**, 99–114 (2014).
- Wessel, P., Smith, W. H. F., Scharroo, R., Luis, J. & Wobbe, F. Generic Mapping Tools: improved version released. *Eos* **94**, 409–410 (2013).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## METHODS

### Satellite source data and implementation of the bloom detection algorithm

We used all Landsat 5 Thematic Mapper (L5 TM) images over study lakes (1984–2012) that covered the five months encompassing summer (June to October or December to April, depending on lake latitude, similar to a global study of satellite-estimated lake temperatures<sup>22</sup>). Given its long-term archive, L5 TM can be used to assess bloom severity in the upper layer of the water column<sup>21,34</sup>. Our analysis used images from the L5 TM top-of-atmosphere (TOA) reflectance image collection in Google Earth Engine<sup>35</sup>, collected originally from the US Geological Survey<sup>36</sup>. The images in this collection represent unitless planetary TOA reflectance ( $\rho_\lambda$ )<sup>37</sup>:

$$\rho_\lambda = \frac{\pi L_\lambda d^2}{\text{ESUN}_\lambda \cos(\theta_s)} \quad (1)$$

where  $L_\lambda$  is the spectral radiance at the sensor's aperture ( $\text{W (m}^2 \text{sr } \mu\text{m)}^{-1}$ ),  $d$  is the Earth–Sun distance (astronomical units),  $\text{ESUN}_\lambda$  is the mean exoatmospheric solar irradiance ( $\text{W (m}^2 \text{ } \mu\text{m)}^{-1}$ ) and  $\theta_s$  is the solar zenith angle (degrees). L5 TM Surface Reflectance products<sup>38</sup> were not available worldwide for the full period at the time of algorithm development<sup>21</sup>.

We implemented a compositing technique to create images covering the whole surface area of each lake. On the basis of the 16-day revisit period of Landsat, we created composites for each lake by stitching together all L5 TM scenes overlapping with the lake during 16-day time intervals. For each year, intervals started on the first day of the first month analysed (either 1 June or 1 December) and ran until 144 days later, or 9 intervals total. Overlapping pixels from multiple scenes were averaged in each final composite. Artefacts of this process were observed for some lakes (for example, horizontal or vertical stripes at scene boundaries), but were not found to substantially affect subsequent analyses.

To mask land within each 16-day composite image, we modified lake polygons from the Global Lakes and Wetlands Database<sup>39</sup>. The original polygons were adjusted manually to fully cover the lake surface area based on images with maximum lake extent over the study period. Within these polygons, we used the Fmask algorithm<sup>40,41</sup>, as implemented in Google Earth Engine<sup>42</sup>, to exclude cloud and cloud shadow pixels, and to identify the water surface area in each composite. We tested the use of other cloud detection algorithms (for example, the Landsat Automatic Cloud Cover Assessment procedure<sup>43</sup>), but Fmask resulted in fewer misclassification errors due to turbid water and haze. We also tested static land cover maps to mask land pixels, but the Fmask water layer better accounted for dynamic changes in lake shorelines over time.

We then applied a bloom detection algorithm based on the near-infrared band, using the shortwave-infrared (SWIR) band to minimize effects of atmospheric interference and a 'greenness' filter to distinguish suspended sediment<sup>21</sup>. The algorithm subtracts the pixel value in the SWIR band (L5 TM band 5, 1.55–1.75  $\mu\text{m}$ ), weighted with an empirical parameter, from the value in the near-infrared band (L5 TM band 4, 0.76–0.90  $\mu\text{m}$ ) as a measure of near-surface bloom intensity:

$$B = F_G(\rho_{B4} - 1.03\rho_{B5}) \quad (2)$$

where  $B$  is the bloom intensity, ranging from 0 to 0.1 where 0.1 generally represents intense near-surface phytoplankton blooms,  $\rho_{B4}$  is L5 TM TOA band 4,  $\rho_{B5}$  is L5 TM TOA band 5 and  $F_G$  represents the greenness filter that masks out pixels below a certain hue ( $H$ ) threshold based on L5 TM TOA in bands 1, 2, and 3:

$$F_G = \begin{cases} 1 & \text{if } H < 1.6 \\ 0 & \text{if } H > 1.6 \end{cases} \quad (3)$$

$$H = \begin{cases} \frac{\rho_{B2} - \rho_{B1}}{\rho_{B3} + \rho_{B2} - 2\rho_{B1}} & \text{if } \rho_{B1} = \min(\rho_{B1}, \rho_{B2}, \rho_{B3}) \\ \frac{\rho_{B3} - \rho_{B2}}{\rho_{B3} + \rho_{B1} - 2\rho_{B2}} + 2 & \text{if } \rho_{B2} = \min(\rho_{B1}, \rho_{B2}, \rho_{B3}) \\ \frac{\rho_{B1} - \rho_{B3}}{\rho_{B2} + \rho_{B1} - 2\rho_{B3}} + 1 & \text{if } \rho_{B3} = \min(\rho_{B1}, \rho_{B2}, \rho_{B3}) \end{cases} \quad (4)$$

Under intense, near-surface algal bloom conditions, backscattering due to phytoplankton abundance dominates the water-leaving radiance in the near-infrared range that is otherwise dampened by pure water absorption<sup>44,45</sup> (Extended Data Figs. 7, 8; see Supplementary Information for additional discussion of algorithm sensitivity to in situ bloom conditions). Although this approach has proven effective in identifying the extent of near-surface intense phytoplankton blooms, we emphasize that the retrieval of concentrations of specific bloom severity metrics (for example, chlorophyll- $a$ ) is beyond the scope of this study.

### Selection of the initial 76 study lakes

We first selected 154 lakes from those included in a study of global lake temperatures<sup>23</sup> with temperature data collected by satellite. The rationale for choosing these lakes was twofold: (1) there was a lower likelihood of persistent cloudiness obscuring images because these lakes had previously been successfully explored using satellite remote sensing and (2) the in situ temperature data for other lakes were not collected in a consistent way that would be representative of the whole lake (for example, some data points were collected at point locations on specific shores of the lake).

After an initial exploration of three randomly selected years of composite images for each lake, we sub-selected 95 lakes for analysis based on the criteria that lake pixels have non-zero bloom intensity values below a threshold of 0.1 in a substantial portion of images. We did this to identify lakes that had ranges of bloom intensity values most similar to Lake Erie, for which the algorithm was originally validated<sup>21</sup>. Compared to the 59 lakes that were not selected, these 95 lakes in general were shallower and had smaller lake volume, suggesting that the approach may have been more applicable for detecting blooms in shallower lakes.

We then performed a literature search to explore whether the observed bloom intensity signal in lakes was likely to be indicative of real phytoplankton blooms (of any type) or a false-positive result. We used ISI Web of Science, Google Scholar and Google Search for the lake name, lake name + "algal bloom" and lake name + "eutrophic". Based on the results of this search, we determined that 78 lakes had either some evidence of phytoplankton blooms (51 out of 78) or no evidence against phytoplankton blooms (27 out of 78), whereas the remaining 17 lakes had strong evidence against the signal representing phytoplankton blooms (for example, in one lake a high bloom intensity signal was erroneously caused by high ice reflectance). Two lakes were also removed from the dataset at this stage based on a lack of L5 TM data during the study period (that is, data were available for only one or two years).

The remaining 76 lakes were selected for further analysis. In these lakes, 54% had support from the literature for evidence of presence of cyanobacteria, with 29% of lakes specifically dominated by *Microcystis* sp. In total, 10,892 composites were compiled from 30,922 L5 TM scenes during the selected months across study lakes, with a median of 139 composites and 283 scenes per lake. The total number of composites per lake ranged from 24 for Lake Edward (43 scenes total) to 250 for Lake Winnipeg (1,771 scenes total). Lakes with a greater number of scenes tended to be in North America, as expected based on historical L5 TM coverage<sup>46</sup>. The number of years with available data per lake ranged from 5 to 28 years with a mean of 21 years (Supplementary Table 1); among lakes with at least 14 years of data, the mean was 22 years per lake.

## Validation of well known spatial gradients in bloom intensity

We further evaluated the proposed approach by comparing geographical regions within lakes with known spatial gradients of bloom intensity. We searched the literature for descriptions of spatial gradients for the 76 study lakes, for example, on the basis of chlorophyll-*a* or phytoplankton biomass observations, and then examined algorithm output values in regions that would be expected to show the largest differences. Using this approach, we identified 48 pairs of regions across 22 lakes (Supplementary Table 2). For instance, in Lake Balaton we identified three regions based on documented chlorophyll-*a* and biomass gradients from the southwest to the northeast<sup>47–49</sup>. The southwest basins are eutrophic to hypertrophic, the northeast basin is mesotrophic and the middle basin is in-between<sup>47</sup>; the expectation is therefore that bloom intensities would be higher in a region in the southwest basin relative to a region in the middle basin, which would then also be higher relative to a region in the northeast basin.

We documented the expected bloom intensity in each region qualitatively (that is, high, medium or low) as well as the strength of the evidence supporting the expected direction of the gradient between regions (that is, strong, medium or weak) (for example, some are based on extensive in situ sampling over many years, whereas others are based on more qualitative inferences). For each of the 48 pairs of regions, we computed the difference in intensity between regions by comparing the mean intensity across all pixels within each region of the lake over the full study period.

The gradient between regions that was inferred using the implemented algorithm was in the correct direction for over three-quarters of region pairs (37 out of 48 pairs). For the Lake Balaton example described earlier, all three comparisons between mean pixel values in the three regions were consistent with the expected sign of the difference in bloom intensity (high versus medium, high versus low, medium versus low).

The previously published evidence was not strong for 10 out of the 11 pairs of regions for which the gradient was in the direction opposite to the expected direction. For example, two pairs were for Songkhla Lake, where the evidence on the spatial gradient came from one publication that was not peer-reviewed and another that was based purely on computational modelling rather than in situ observations<sup>50,51</sup>. Because the vast majority of the discrepancies in identifying spatial gradients were for region pairs with weaker support in the literature, results indicate that the implemented algorithm is able to identify well established spatial gradients in bloom intensity across a variety of lakes.

We note that the qualitative approach used here represents a first step towards understanding how global patterns in bloom intensity may be measured using satellite remote sensing. Although the algorithm used here was validated quantitatively for data from only Lake Erie, the results of our qualitative approach together with atmospheric radiative transfer simulations (see Supplementary Information), indicate that the algorithm provides a useful signal of bloom intensity for the lakes analysed in this study. Additional validation may be needed for other applications.

## Generation of bloom intensity time series and trends

To generate long-term time series of summertime maximum bloom intensities for each lake, we summed the algorithm output values over the whole lake for each composite, as a measure of bloom intensity. This approach assumes that the algorithm output value is correlated with measures such as near-surface chlorophyll-*a* or phytoplankton biomass, which is supported by validation based on well known spatial gradients, as described above. We took the largest composite bloom intensity each year as an estimate of summertime maximum bloom intensity, similar to other studies<sup>21,52</sup>. This approach focuses on the relative spatial intensity of phytoplankton blooms over time, and therefore minimizes the impact of noise in L5 TM images over bodies of water<sup>53</sup>.

We removed estimates from each time series when the observed lake surface area was less than 80% of maximum, or less than the mean minus one s.d. of the whole surface area of the lake for the time series, whichever was lower. These guidelines were determined heuristically, based on removing composites that visually had large portions of the lake surface area unavailable due either to missing L5 TM scenes or high cloud cover. We also removed estimates for years in which there were fewer than 3 composites due to missing data, as these were expected to be less representative of the summertime maximum. Main findings were not found to be sensitive to minor variations in these thresholds.

We further adjusted for variations in observed lake surface due to clouds or missing L5 TM scenes by dividing the annual bloom intensity estimates by the observed surface area of each lake in each year. This had the benefit of correctly adjusting observed bloom severity trends for lakes where the water surface area changed substantially over time. For example, for a subset of lakes that have dried up during the study period (for example, Lake Urmia or the Aral Sea), a decline in water surface area would otherwise be incorrectly observed as a decline in bloom intensity. Normalizing by water surface area more accurately reflects the true bloom conditions in those lakes over time. Although, in principle, this could also make blooms of constant severity in an otherwise shrinking lake appear to have an increasing trend, this scenario was not found among study lakes with declining water surface area. Given the highly varying local conditions for study lakes with respect to bloom intensity and water surface area trends with time, normalizing by water surface area provided the best approach overall for accounting for variations due to clouds and missing L5 TM scenes.

Finally, to compare data across different lakes, we normalized the annual time series of the peak bloom intensity for each lake by its own long-term mean and s.d. values, creating bloom intensity z-scores. This is similar to other studies that have treated historical bloom data from remote sensing<sup>54</sup>, tracked long-term trends in cyanobacteria<sup>6</sup> or estimated long-term trends of other parameters in lakes<sup>27</sup>.

## Evaluation of bloom intensity time series and trends

To further evaluate the implemented approach, we compared the temporal evolution of peak bloom intensity in well studied lakes to those described in existing literature, and also compared bloom intensity trends overall to trends in the SWIR TOA reflectance.

From each time series of normalized annual peak bloom intensity, we tested for the presence of monotonic time trends using the *S* statistic from the Mann–Kendall trend test<sup>55</sup> and estimated the magnitude of temporal trends using Thiel–Sen’s slope<sup>56</sup>. These tests are both non-parametric procedures that are known to be more robust to outliers and more accurate for skewed or heteroskedastic data<sup>57</sup> and have been used widely for assessing temporal trends in limnologic studies and those evaluating trends in phytoplankton blooms specifically<sup>6,31,58,59</sup>. Trend analyses over the whole study period were performed for all lakes ( $n = 76$ ).

For lakes in which large changes in bloom intensity have been documented for the study period, the data used here accurately matched both the direction and timing of changes described in the literature. For example, substantial improvements in bloom conditions have been reported for Lake Balaton<sup>60,61</sup>, Clear Lake<sup>62,63</sup> and Lake Simcoe<sup>16,64</sup>, and all showed statistically significant ( $P < 0.1$ ) decreases in peak bloom intensity in the data presented here over the same timeframes as described in previous studies. The improvement in water quality of Lake Balaton that occurred in the 1990s, which coincided with sewage controls and a decline in agriculture<sup>61</sup>, was reproduced correctly in the time series (Fig. 4a). Clear Lake experienced a similar decline in bloom intensity in the 1990s—probably due to the trophic cascade stemming from drought<sup>63</sup>—that was also correctly reproduced (Fig. 4b). Furthermore, for Lake Simcoe, an improvement in water quality that occurred soon after 1995, which coincided with a widespread invasion of zebra

mussels<sup>16,65,66</sup>, was also successfully reproduced (Fig. 4d). The developed data similarly reproduced histories for lakes with documented increases in bloom intensity during the study period, such as Lake Winnipeg<sup>67</sup> and Lake Baikal<sup>4,58</sup>, and captured documented temporal patterns of decreasing and increasing bloom intensities in ecosystems as diverse as Lake Erie<sup>68</sup> and Tsimlyansk Reservoir<sup>69</sup> (Extended Data Fig. 9).

We further evaluated the performance of the approach for the detection of false-positive (that is, high derived bloom intensity for low bloom intensity spectra) and false-negative (that is, low derived bloom intensity for high bloom intensity spectra) results using the atmospheric radiative transfer simulations described in the Supplementary Information. The analyses demonstrated the robustness of the algorithm in limiting false-positive results—that is, correctly identifying instances in which no blooms were present. However, the analyses also identified that aerosol optical thickness (AOT) has an effect on the incidence of false-negative results whereby a higher AOT (for example, hazy conditions) resulted in an increased likelihood that high bloom intensity events were missing.

Because SWIR can be used as a proxy for AOT, we next evaluated whether derived bloom intensity trends in study lakes could erroneously be due to trends in aerosol conditions that affect the likelihood of false-negative results. This was accomplished by comparing peak bloom intensity trends to trends in the SWIR TOA reflectance. For five lakes with statistically significant trends ( $P < 0.1$ ) in both peak bloom intensity and SWIR TOA reflectance, we found that observed bloom intensity trends coincided with trends in SWIR TOA reflectance, increasing the risk that apparent trends in bloom intensity could be due to a change in the likelihood of false-negative results. For these lakes, peak bloom intensity and SWIR trends were in opposite directions (that is, increasing SWIR resulting in an increased likelihood of false-negative results consistent with a decreased bloom intensity trend, and vice versa for decreasing SWIR). Decreasing bloom intensity trends were observed in four of the five lakes (Chao, Gaoyou, Taihu and Sarykamshskoye), with an increasing bloom intensity trend observed in the fifth (Kremenshugskoye). Increasing SWIR trends in the four lakes were consistent with trends in aerosols in Eastern China and central Asia over the study period<sup>70</sup>, whereas a reduction in aerosols in Eastern Europe<sup>70</sup> was consistent with the latter. Because we could not confirm whether or not the observed bloom intensity trends were in fact attributable to trends in SWIR, to be conservative we removed these five lakes from the subsequent analysis. This resulted in a final set of  $n = 71$  lakes for further analysis. For trends estimated over shorter periods (1984–1997 and 1998–2012), slightly fewer lakes were used ( $n = 66$  and  $n = 68$ , respectively) because at least two years of observations per period were required to compute trends. A subset of lakes with at least 14 years of data ( $n = 49$ ) was also used to explore temporal patterns in peak bloom intensity.

Beyond this analysis of SWIR trends, we found no other evidence that any potential misclassification of bloom intensity trends occurred. To assess whether our approach could have been incorrectly measuring trends in other environmental variables, we explored historical patterns of potential confounders that have been documented in the literature. Trends in Secchi depth did not match the observed bloom intensity trends (for example, in Lake Simcoe<sup>71</sup>, Great Salt Lake<sup>72</sup> or Lake Okeechobee<sup>73</sup>), indicating a lack of evidence to suggest that the implemented algorithm was potentially measuring other constituents of water quality. Nor did we find any evidence that global changes in atmospheric constituents, such as aerosols, dust and water vapour, could explain the overall geographical pattern of bloom observations (except in the five aforementioned lakes), because such constituents have a spatial coherence at large regional scales<sup>70,74–76</sup> whereas the observed lake trends were highly spatially heterogeneous (Fig. 1). For individual lakes, bloom intensity trends also did not track well with observed trends in submerged aquatic vegetation (for example, Lake Okeechobee<sup>73</sup>), gypsum (for example, Salton Sea<sup>77</sup>) or cloud cover (for example, Lake Nicaragua<sup>78</sup>). Taken together, this suggested that our findings about

the global proportion of lakes with increasing bloom intensity trends were likely to be robust.

### Characteristics of the final 71 study lakes

The 71 study lakes (Supplementary Table 1) spanned a wide range of surface areas (158 to 67,052 km<sup>2</sup>) and maximum depth (2 to 1,637 m) comparable to previous global studies of lakes<sup>29</sup>. Of the 49 lakes with at least 14 years of bloom data, a large majority warmed over the study period (88%) (Extended Data Fig. 6), with the temperature trend ranging from  $-1.40$  °C per decade (that is, cooling) to  $0.93$  °C per decade. Most of these 49 lakes also experienced an increase in annual precipitation (61%, ranging from  $-49$  to  $173$  mm per decade) while close to half of the lakes experienced an increase in fertilizer application rate (49%, ranging from  $-1.47$  to  $2.41$  Mg N km<sup>-2</sup> per decade) (Supplementary Table 1).

### Categorization of lakes by prototypical historical pathway

To bin the lakes by prototypical historical pathway, we fit a linear model with time for each lake time series using ordinary least squares regression:

$$y = \beta_1 t + \beta_0 \quad (5)$$

where  $y$  represents the normalized maximum summertime bloom intensity,  $t$  represents the year of the observation, and  $\beta_1$  and  $\beta_0$  are the fitted model parameters. Bloom intensity values from individual lake pixels ( $B$  from Eq. (2)) were summed for each image composite, and the maximum summed bloom intensity for each year was used to create the time series  $y$  after correcting for missing data, subtracting the long-term mean and dividing by the long-term s.d. Lakes for which the linear term was statistically significant ( $P < 0.1$ ) were categorized as sustained improvement if the peak bloom intensity trend decreased with time ( $\beta_1 < 0$ ) and deterioration if the peak bloom intensity trend increased with time ( $\beta_1 > 0$ ).

For the remaining lakes, we fit a quadratic model to each lake time series:

$$y = \beta_2 t^2 + \beta_1 t + \beta_0 \quad (6)$$

where a third term is added indicating a change in peak bloom intensity with  $t^2$ . Lakes for which the quadratic term was statistically significant ( $P < 0.1$ ) and  $\beta_2 > 0$  were categorized as improvement then deterioration. The remaining lakes were categorized as no significant trend.

This approach used both simple monotonic trends with time as well as an assessment of the degree of improvement and deterioration to categorize bloom intensity trends. Categorization of lakes (eutrophication, restoration or no change) on the basis of simple changes in lake parameters (increasing, decreasing or no consistent change, respectively) has been used previously to understand long-term trends<sup>6</sup>, as have comparisons of multiple measurements over time to assess the balance between historical deterioration and improvement<sup>18</sup>.

### Data availability

The Landsat 5 Thematic Mapper imagery used in this study is available from the US Geological Survey (<http://earthexplorer.usgs.gov>) and through Google Earth Engine (<https://earthengine.google.com>). The bloom intensity trend estimates, historical pathway categories and environmental driver variables generated for each lake and analysed in this study are provided in Supplementary Table 1. The temperature, precipitation, fertilizer use and lake geomorphological data supporting the findings of this study are publicly available<sup>23,79,80</sup> (see ‘Environmental driver, watershed, and geomorphological characteristic data sets’ in the Supplementary Information).



## Code availability

Google Earth Engine's web interface allows the bloom detection algorithm<sup>21</sup> to be applied on any Landsat 5 Thematic Mapper images. Access will be provided upon request.

34. Tebbis, E. J., Remedios, J. J. & Harper, D. M. Remote sensing of chlorophyll-a as a measure of cyanobacterial biomass in Lake Bogoria, a hypertrophic, saline-alkaline, flamingo lake, using Landsat ETM+. *Remote Sens. Environ.* **135**, 92–106 (2013).
35. Google. Earth Engine. <https://earthengine.google.com/> (2016).
36. USGS. Landsat Missions. <http://landsat.usgs.gov/> (2016).
37. Chander, G., Markham, B. L. & Helder, D. L. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. *Remote Sens. Environ.* **113**, 893–903 (2009).
38. USGS. Landsat Surface Reflectance Level-2 Data Products. <https://landsat.usgs.gov/landsat-surface-reflectance-data-products> (2017).
39. Lehner, B. & Döll, P. Development and validation of a global database of lakes, reservoirs and wetlands. *J. Hydrol.* **296**, 1–22 (2004).
40. Zhu, Z. & Woodcock, C. E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **118**, 83–94 (2012).
41. Zhu, Z., Wang, S. & Woodcock, C. E. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4-7, 8, and Sentinel 2 images. *Remote Sens. Environ.* **159**, 269–277 (2015).
42. Erickson, T. A. Earth Engine Data Catalog: USGS Landsat 5 TOA Reflectance (Orthorectified) with Fmask. [https://code.earthengine.google.com/dataset/LANDSAT/LT5\\_L1T\\_TOA\\_FMASK](https://code.earthengine.google.com/dataset/LANDSAT/LT5_L1T_TOA_FMASK) (2016).
43. Irish, R. R. Landsat 7 automatic cloud cover assessment. In *Proc. SPIE 4049, Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI* 348–355 (2000).
44. Moore, T. S. et al. Bio-optical properties of cyanobacteria blooms in western Lake Erie. *Front. Mar. Sci.* **4**, 300 (2017).
45. Gower, J., King, S., Borstad, G. & Brown, L. Use of the 709 nm band of MERIS to detect intense plankton blooms and other conditions in coastal waters. *ESA J.* **1161**, 365–368 (2005).
46. Goward, S. et al. Historical record of Landsat global coverage: mission operations, NSLRSDA, and international cooperator stations. *Photogramm. Eng. Remote Sensing* **72**, 1155–1169 (2006).
47. Palmer, S. C. J. et al. Validation of Envisat MERIS algorithms for chlorophyll retrieval in a large, turbid and optically-complex shallow lake. *Remote Sens. Environ.* **157**, 158–169 (2015).
48. Pálffy, K., Présing, M. & Vörös, L. Diversity patterns of trait-based phytoplankton functional groups in two basins of a large, shallow lake (Lake Balaton, Hungary) with different trophic state. *Aquat. Ecol.* **47**, 195–210 (2013).
49. Hajnal, É. & Padisák, J. Analysis of long-term ecological status of Lake Balaton based on the ALMOBAL phytoplankton database. *Hydrobiologia* **599**, 227–237 (2008).
50. Chesoh, S., Lim, A. & Tongkumchum, P. Trend of water quality and model for forecasting eutrophication occurrence in Songkhla Lake, Thailand. In *Proc. Taal2007: The 12th World Lake Conference* 834–839 (2008).
51. Suwanichcharoen, S. & Liengcharernsit, W. Development of phytoplankton model with application to Songkhla Lake, Thailand. *Low. Technol. Int.* **14**, 50–59 (2012).
52. Stumpf, R. P., Wynne, T. T., Baker, D. B. & Fahnenstiel, G. L. Interannual variability of cyanobacterial blooms in Lake Erie. *PLoS ONE* **7**, e42444 (2012).
53. Pahlevan, N., Balasubramanian, S. V., Sarkar, S. & Franz, B. A. Toward long-term aquatic science products from heritage Landsat missions. *Remote Sens.* **10**, 1337 (2018).
54. Palmer, S. C. J. et al. Satellite remote sensing of phytoplankton phenology in Lake Balaton using 10 years of MERIS observations. *Remote Sens. Environ.* **158**, 441–452 (2015).
55. Mann, H. B. Nonparametric tests against trend. *Econometrica* **13**, 245–259 (1945).
56. Sen, P. K. Estimates of the regression coefficient based on Kendall's tau. *J. Am. Stat. Assoc.* **63**, 1379–1389 (1968).
57. Rousseeuw, P. J. & Leroy, A. M. *Robust Regression and Outlier Detection* (John Wiley & Sons, 1987).
58. Izmeteva, L. R. et al. Lake-wide physical and biological trends associated with warming in Lake Baikal. *J. Great Lakes Res.* **42**, 6–17 (2016).
59. Gobler, C. J. et al. Ocean warming since 1982 has expanded the niche of toxic algal blooms in the North Atlantic and North Pacific oceans. *Proc. Natl Acad. Sci. USA* **114**, 4975–4980 (2017).
60. Padisák, J. & Koncos, L. Trend and noise: long-term changes of phytoplankton in the Keszthely Basin of Lake Balaton, Hungary. *Int. Assoc. Theor. Appl. Limnol.* **28**, 194–203 (2002).
61. Tátrai, I., Istvánovics, V., Tóth, L.-G. & Kóbor, I. Management measures and long-term water quality changes in Lake Balaton (Hungary). *Fundam. Appl. Limnol.* **172**, 1–11 (2008).
62. Mioni, C., Kudela, R., Baxa, D. & Sullivan, M. *Harmful Cyanobacteria Blooms and their Toxins in Clear Lake and the Sacramento-San Joaquin Delta (California)* (Central Valley Regional Water Quality Control Board, 2011).
63. Winder, M., Reuter, J. & Schladow, G. *Clear Lake Historical Data Analysis* (Univ. California, Davis, 2010).
64. North, R. L. et al. The state of Lake Simcoe (Ontario, Canada): the effects of multiple stressors on phosphorus and oxygen dynamics. *Inland Waters* **3**, 51–74 (2013).
65. Evans, D. O., Skinner, A. J., Allen, R. & McMurtry, M. J. Invasion of zebra mussel, *Dreissena polymorpha*, in Lake Simcoe. *J. Great Lakes Res.* **37**, 36–45 (2011).
66. Baranowska, K. A., North, R. L., Winter, J. G. & Dillon, P. J. Long-term seasonal effects of dreissenid mussels on phytoplankton in Lake Simcoe, Ontario, Canada. *Inland Waters* **3**, 285–296 (2013).
67. Schindler, D. W., Hecky, R. E. & McCullough, G. K. The rapid eutrophication of Lake Winnipeg: greening under global change. *J. Great Lakes Res.* **38**, 6–13 (2012).
68. Allinger, L. & Reavie, E. The ecological history of Lake Erie as recorded by the phytoplankton community. *J. Great Lakes Res.* **39**, 365–382 (2013).
69. Nilson, E. *Investigating Potential Agricultural-related Causes of Eutrophication in the Tsimlyansk Reservoir through GIS and Remote Sensing*. MSc thesis, Central European Univ. (2014).
70. Pozzer, A. et al. AOD trends during 2001–2010 from observations and model simulations. *Atmos. Chem. Phys.* **15**, 5521–5535 (2015).
71. Guan, X., Li, J. & Booty, W. G. Monitoring Lake Simcoe water clarity using Landsat-5 TM images. *Water Resour. Manage.* **25**, 2015–2033 (2011).
72. Belovsky, G. E. et al. The Great Salt Lake Ecosystem (Utah, USA): long term data and a structural equation approach. *Ecosphere* **2**, art33 (2011).
73. Havens, K. et al. Extreme weather events and climate variability provide a lens to how shallow lakes may respond to climate change. *Water* **8**, 229 (2016).
74. Chin, M. et al. Multi-decadal aerosol variations from 1980 to 2009: a perspective from observations and a global model. *Atmos. Chem. Phys.* **14**, 3657–3690 (2014).
75. Ginoux, P., Prospero, J. M., Gill, T. E., Hsu, N. C. & Zhao, M. Global-scale attribution of anthropogenic and natural dust sources and their emission rates based on MODIS deep blue aerosol products. *Rev. Geophys.* **50**, RG3005 (2012).
76. Wang, J., Dai, A. & Mears, C. Global water vapor trend from 1988 to 2011 and its diurnal asymmetry based on GPS, radiosonde, and microwave satellite measurements. *J. Clim.* **29**, 5205–5222 (2016).
77. Tiffany, M. A., Ustin, S. L. & Hurlbert, S. H. Sulfide eruptions and gypsum blooms in the Salton Sea as detected by satellite imagery, 1979–2006. *Lake Reserv. Manage.* **23**, 637–652 (2007).
78. Chang, N.-B., Bai, K. & Chen, C.-F. Smart information reconstruction via time-space-spectrum continuum for cloud removal in satellite images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **8**, 1898–1912 (2015).
79. Wei, Y. et al. NACP MsTMIP: Global and North American Driver Data for Multi-Model Intercomparison. <https://doi.org/10.3334/ORNLDAA/1220> (ORNL DAAC, 2014).
80. Center for International Earth Science Information Network - CIESIN - Columbia University. Gridded Population of the World, Version 4 (GPWv4): Population Density. <https://doi.org/10.7927/H4NP22DQ> (NASA Socioeconomic Data and Applications Center (SEDAC), 2016).

**Acknowledgements** We thank T. Ballard and D. Del Giudice for discussions and input, Y. Fang and E. Sinha for help with obtaining and processing environmental driver datasets, as well as N. Gorelick and T. A. Erickson for guidance with Google Earth Engine. This research was supported by the National Science Foundation (NSF) under grant 1313897. Additional support was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) under a Postgraduate Scholarship-Doctoral award (PGSD3-438855-2013), by a 2015 Google Earth Engine Research Award, by NASA ROSES grant NNX16AI16G and by USGS Landsat Science Team Award 140G0118C0011. CRU-NCEP precipitation was provided by the Multi-scale synthesis and Terrestrial Model Intercomparison Project (MsTMIP; <http://nacp.ornl.gov/MsTMIP.shtml>). Funding for the MsTMIP activity was provided through NASA grant NNX10AG01A. Data management support for preparing, documenting and distributing model driver data was performed by the Modeling and Synthesis Thematic Data Center at Oak Ridge National Laboratory (ORNL; <http://nacp.ornl.gov>), with funding from NASA grant NNN10AN681.

**Author contributions** J.C.H. and A.M.M. designed the research and analysed the results. J.C.H. and A.M.M. wrote the manuscript with input from N.P. J.C.H. performed the majority of the computations with input from A.M.M. N.P. performed the MODTRAN simulations, analysed the MODTRAN results and wrote the corresponding sections of the Methods.

**Competing interests** The authors declare no competing interests.

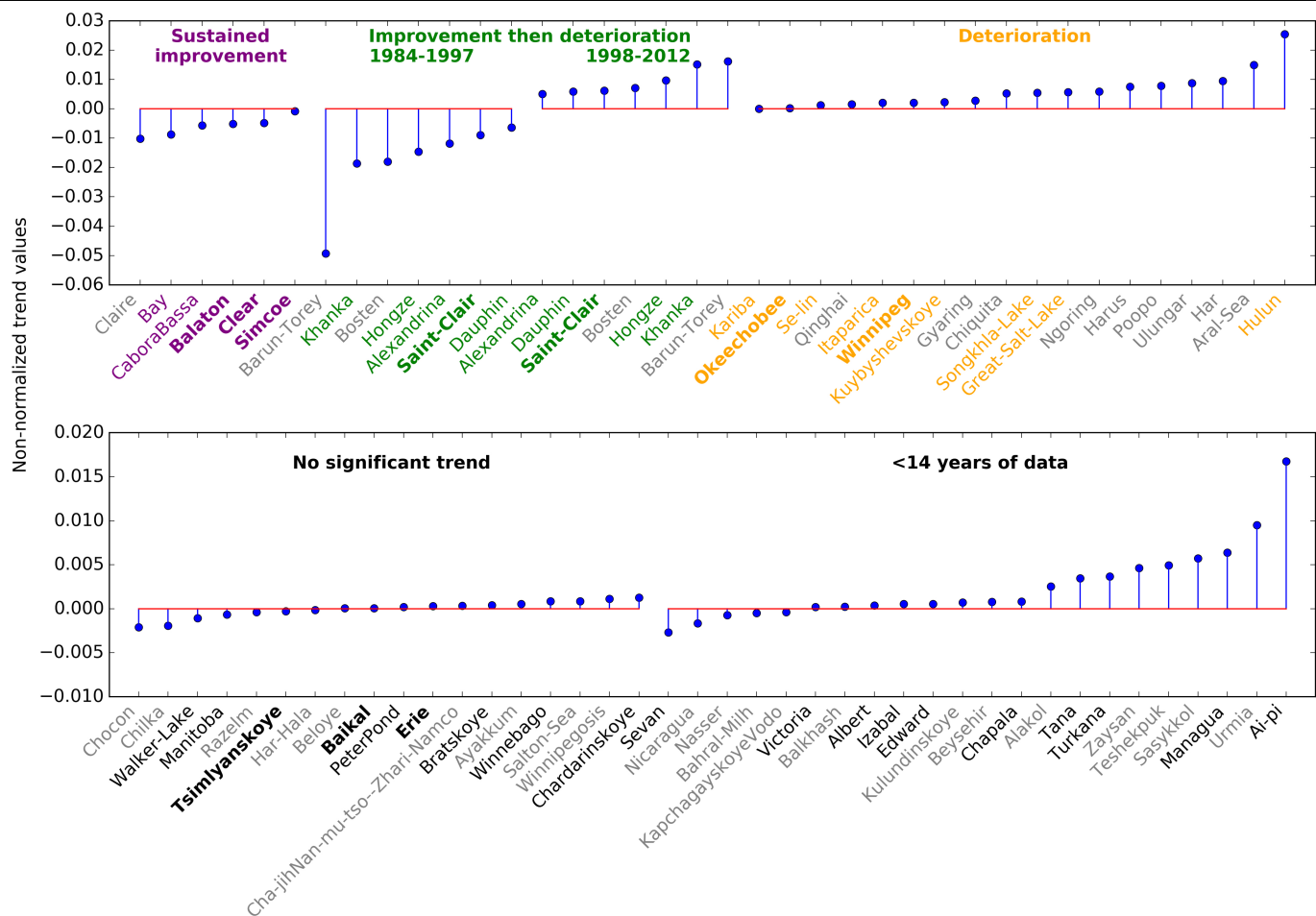
### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1648-7>.

**Correspondence and requests for materials** should be addressed to J.C.H. or A.M.M.

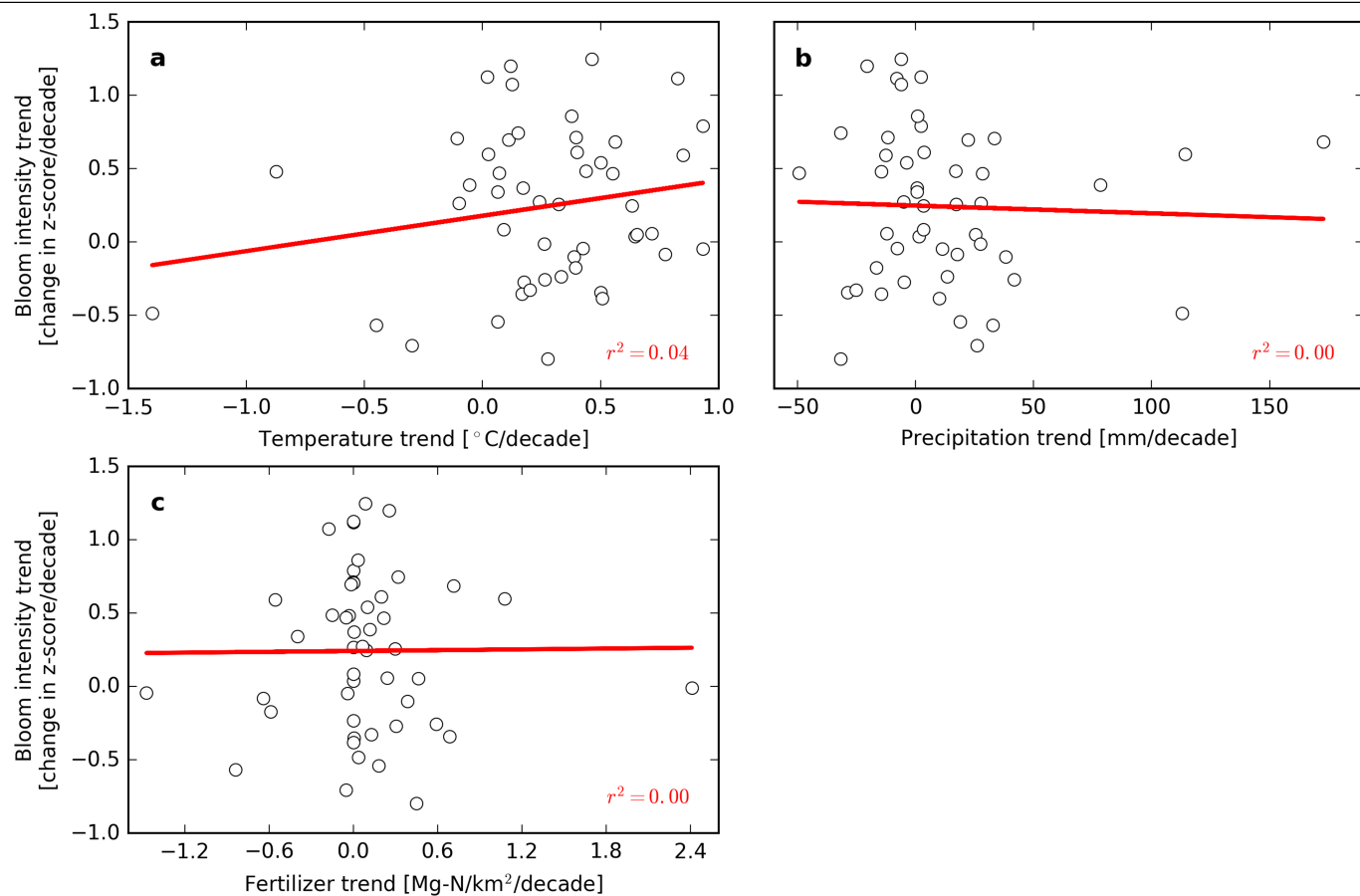
**Peer review information** *Nature* thanks Xi Chen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



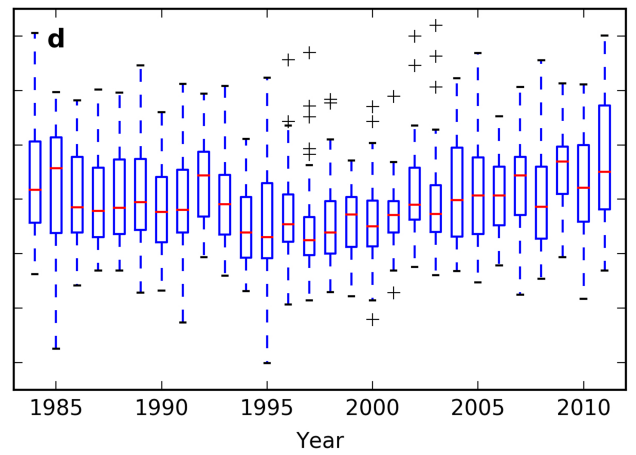
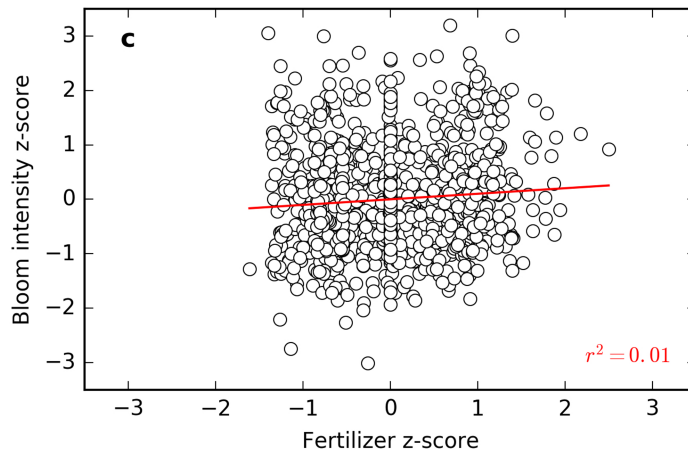
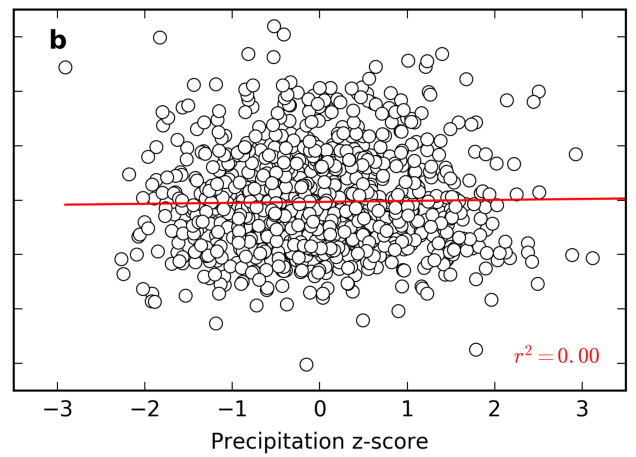
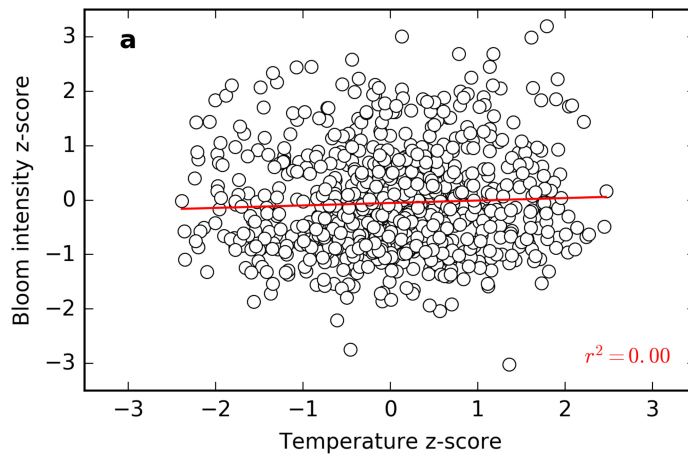
**Extended Data Fig. 1 | Lakes with evidence of cyanobacteria and with well documented evidence of major ecological changes show trends of a similar magnitude to the trends in other lakes.** Lake names in colour indicate that there is evidence of cyanobacteria in that lake; bold lake names indicate that there is evidence of major ecological changes. The y axis shows temporal trends in peak bloom intensity before normalization for all 71 study lakes, categorized

by historical pathway. These temporal trends are the Thiel-Sen's slope values calculated using the maximum summertime lake-wide bloom intensity time series for each lake. Trends for lakes of the 'improvement then deterioration' pathway are separated into trends for 1984-1997 and 1998-2012 to show trend values in each sub-period separately.



**Extended Data Fig. 2 | Low correlations between trends in bloom intensity and environmental factors.** a–c, Scatter plots of the trend in bloom intensity compared with the trends in temperature (a), total precipitation (b) and

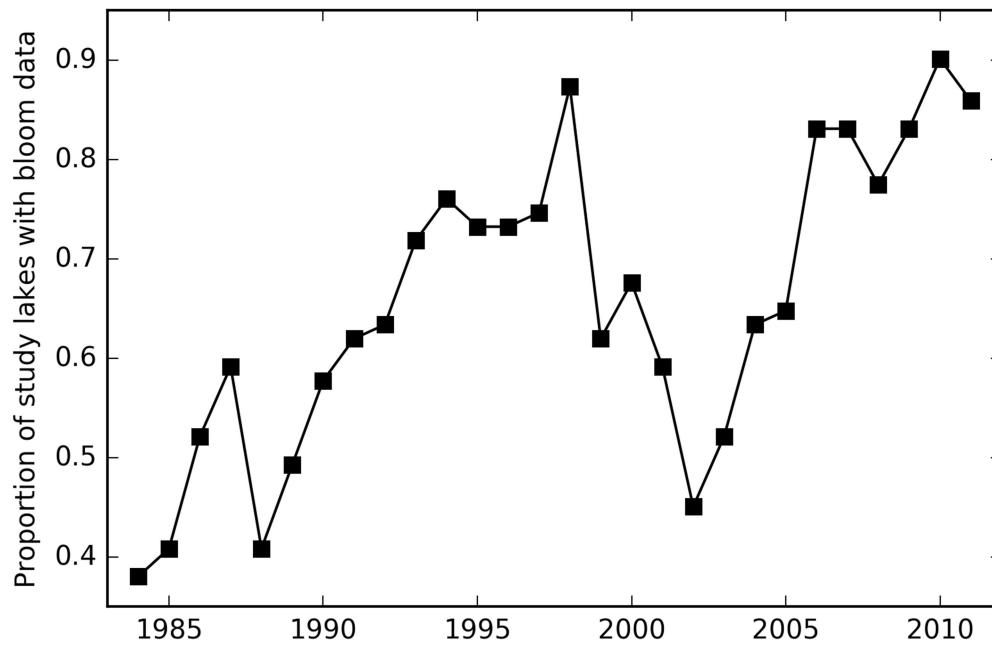
fertilizer application (c) for study lakes with at least 14 years of data ( $n = 49$ ). Each circle represents one lake. Red lines indicate the linear fit of the white circles.



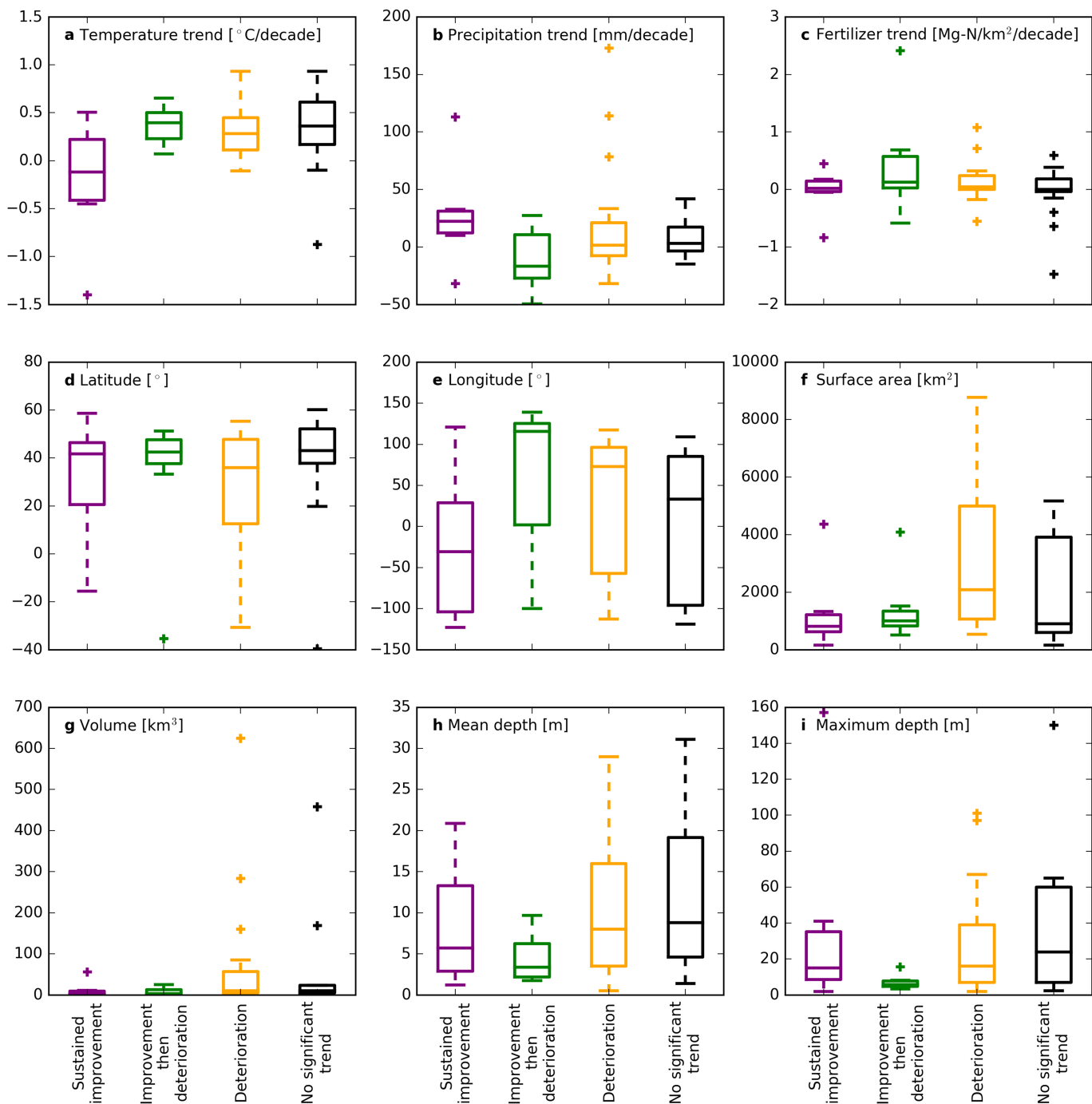
**Extended Data Fig. 3 | No relationship is observed between bloom intensity and environmental factors collected from all lakes. a–c.** Scatter plots of bloom intensity z-score compared with temperature (a;  $n = 784$ ), precipitation (b;  $n = 936$ ) and fertilizer (c;  $n = 980$ ) z-scores. Each circle represents one year for one lake. The z-score of each lake variable is calculated using the mean and s.d. of its own time series. Red lines indicate the linear fit of the white circles. **d.** Box

plots of bloom intensity z-score ( $n = 980$  total). Each box plot shows the distribution of z-scores for all lakes with available data each year. Each box extends from the first to the third quartile values, with a line at the median. The whiskers extend to  $1.5 \times$  the interquartile range from the edges of the box. The plus symbols show outlier values past the end of the whiskers.

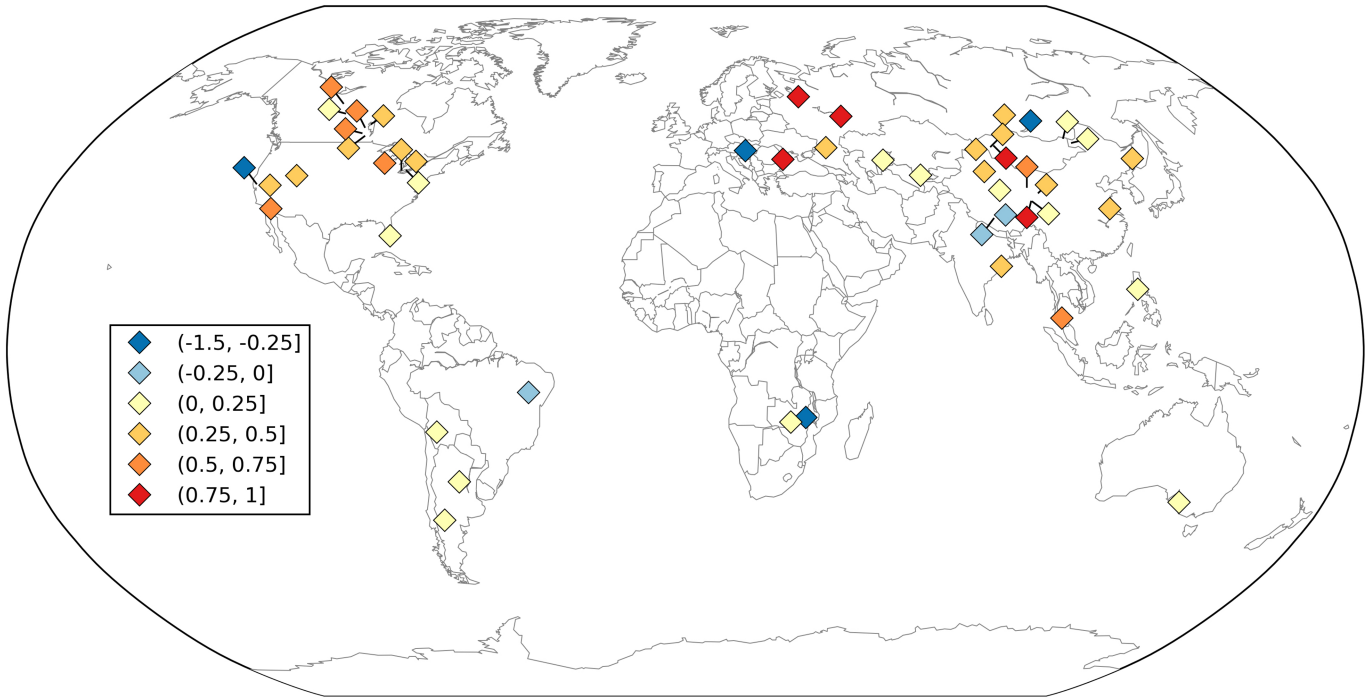




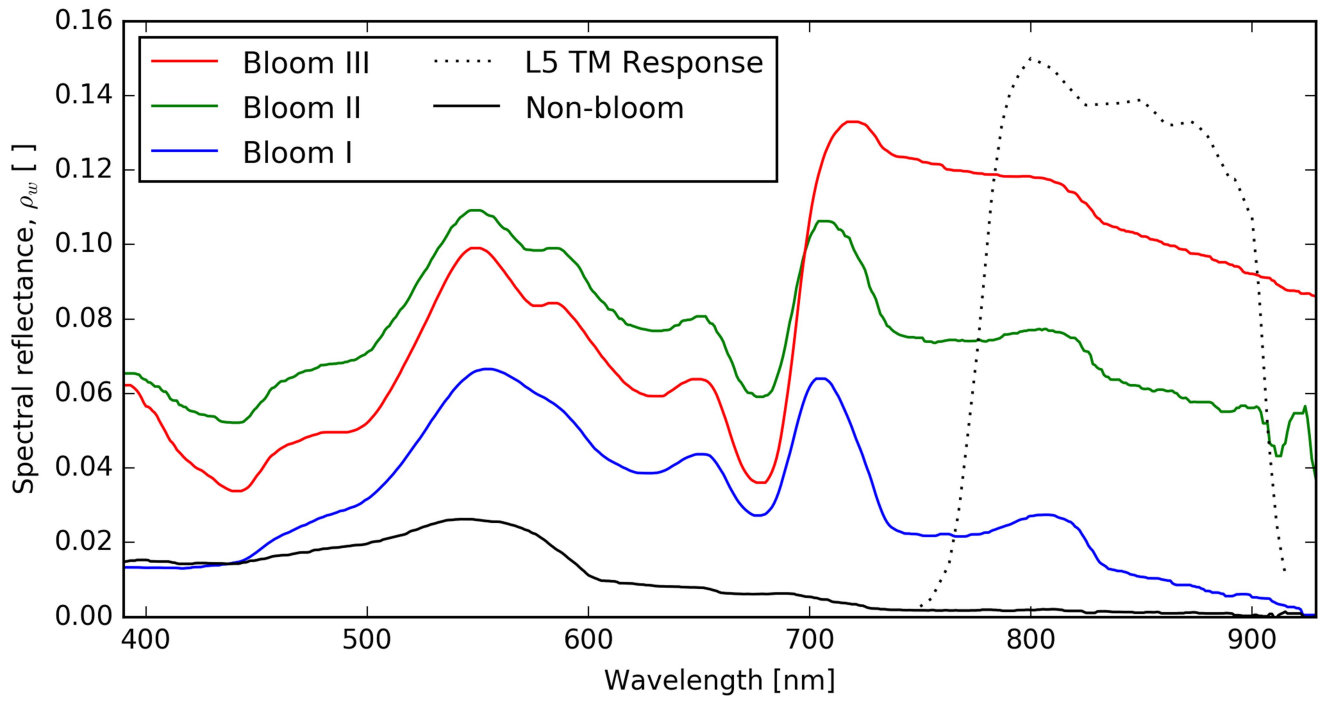
**Extended Data Fig. 4 | Availability of bloom intensity data during the study period.** Number of lakes with a bloom intensity observation after correction for clouds and number of composites (see Methods) divided by the total number of study lakes ( $n = 71$ ) for each year.



**Extended Data Fig. 5 | Distributions of lake variables by historical pathway.** **a–c**, Distributions of environmental drivers. **d–i**, Distributions of geomorphological factors. The data in **a** are equivalent to those in Fig. 3.



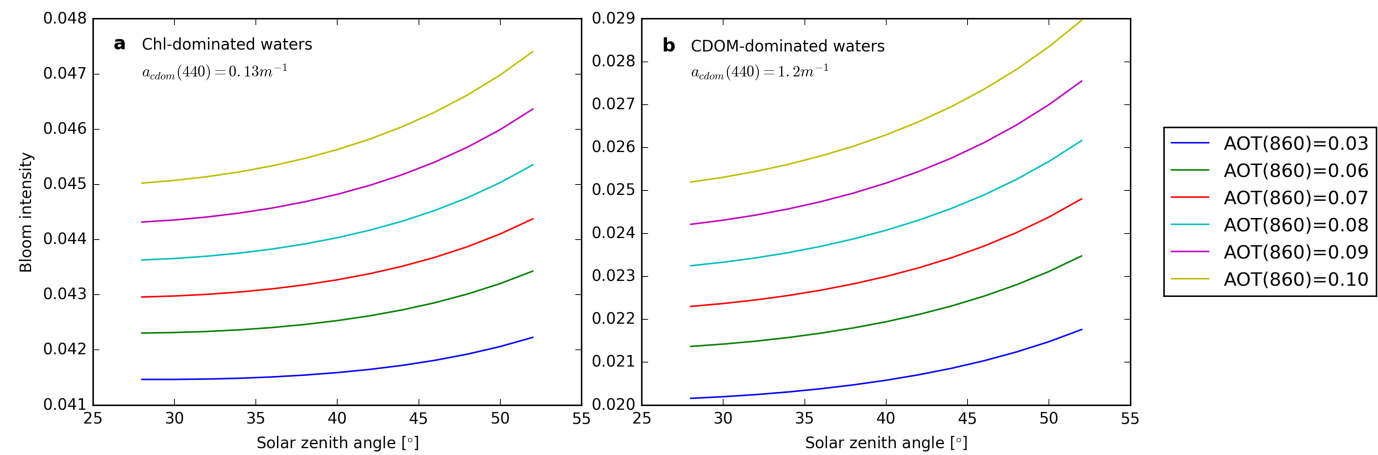
**Extended Data Fig. 6 | Global distribution of trends in lake temperature.** For the lakes with at least 14 years of bloom data ( $n = 49$ ), the maps show the temporal trend in lake surface water temperature ( $^{\circ}\text{C}$  per decade). The base map was generated using Generic Mapping Tools<sup>33</sup>.



**Extended Data Fig. 7 | Spectral reflectance curves used in simulations to test algorithm sensitivity.** Spectral reflectance curves ( $\rho_w[-]$ ) associated with three phytoplankton bloom concentrations and one non-bloom water condition measured in Lake Erie are shown<sup>44</sup>. Blooms I, II and III correspond to near-surface chlorophyll-*a* concentrations of  $100.1 \text{ mg m}^{-3}$ ,  $143.7 \text{ mg m}^{-3}$  and  $106.3 \text{ mg m}^{-3}$ , respectively, and total suspended solid concentrations of  $30.1 \text{ g m}^{-3}$ ,  $20.0 \text{ g m}^{-3}$

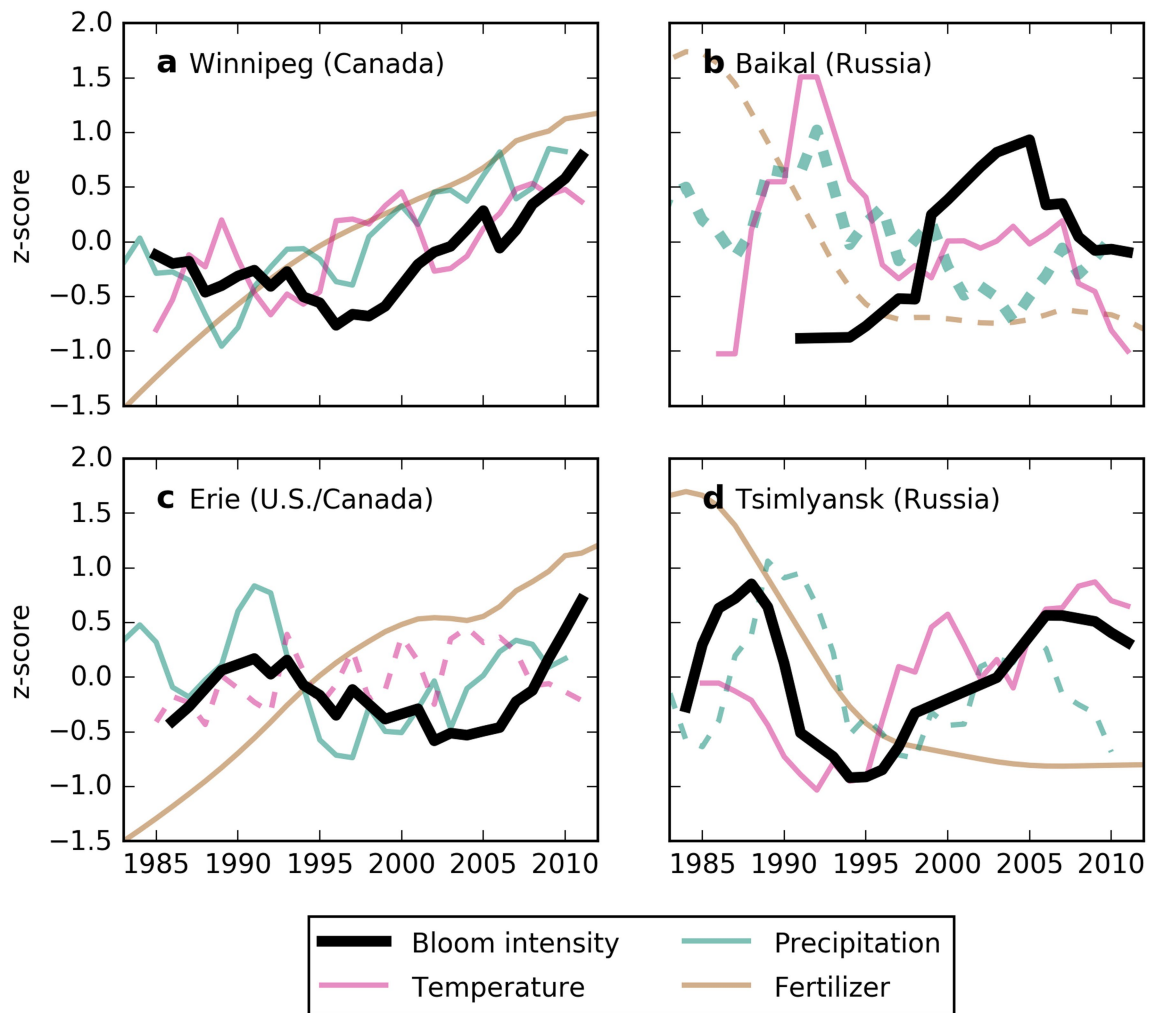
and  $22.7 \text{ g m}^{-3}$ , respectively. The non-bloom curve corresponds to chlorophyll-*a* and total suspended solid concentrations of  $5.8 \text{ mg m}^{-3}$  and  $1.8 \text{ g m}^{-3}$ , respectively. The normalized spectral response of the near-infrared channel of L5 TM is also shown. The spectra used in the sensitivity analyses demonstrate the robustness of the bloom intensity measure used in this study (Eq. (2)).





**Extended Data Fig. 8 | Observed bloom intensity shows minimal sensitivity to changes in aerosol optical thickness (AOT) or to changes in the solar zenith angle that would result from a change in Landsat 5 orbit. a, b,** The sensitivity of derived bloom intensity varies on the order of 0.001 for waters dominated by chlorophyll (Chl) (a) and coloured dissolved organic matter

(CDOM) (b) for changes in solar zenith angle that would be expected owing to a change in satellite orbit. The simulated variation due to solar zenith angle is even smaller for coarse aerosol types (that is, smaller values of AOT). The environments in a and b correspond to bloom III and bloom I, respectively, in Extended Data Fig. 7.



**Extended Data Fig. 9 | Historical bloom intensity patterns for four additional lakes with well documented temporal trends.** Graphs as in Fig. 4 for four additional lakes that had well documented temporal trends. Panels show five-year moving averages of normalized bloom intensity, summer lake

temperatures, and total precipitation and fertilizer application rate over each lake's watershed. Thicker temperature, precipitation and fertilizer lines indicate that the Pearson correlation coefficient with bloom intensity is significant ( $P < 0.1$ ). Dashed lines indicate anti-correlations.

# Arthropod decline in grasslands and forests is associated with landscape-level drivers

<https://doi.org/10.1038/s41586-019-1684-3>

Received: 8 February 2019

Accepted: 16 September 2019

Published online: 30 October 2019

Sebastian Seibold<sup>1,2\*</sup>, Martin M. Gossner<sup>3</sup>, Nadja K. Simons<sup>1,4</sup>, Nico Blüthgen<sup>4</sup>, Jörg Müller<sup>2,5</sup>, Didem Ambarlı<sup>1,6</sup>, Christian Ammer<sup>7</sup>, Jürgen Bauhus<sup>8</sup>, Markus Fischer<sup>9</sup>, Jan C. Habel<sup>1,10</sup>, Karl Eduard Linsenmair<sup>11</sup>, Thomas Nauss<sup>12</sup>, Caterina Penone<sup>9</sup>, Daniel Prati<sup>9</sup>, Peter Schall<sup>7</sup>, Ernst-Detlef Schulze<sup>13</sup>, Juliane Vogt<sup>1</sup>, Stephan Wöllauer<sup>12</sup> & Wolfgang W. Weisser<sup>1</sup>

Recent reports of local extinctions of arthropod species<sup>1</sup>, and of massive declines in arthropod biomass<sup>2</sup>, point to land-use intensification as a major driver of decreasing biodiversity. However, to our knowledge, there are no multisite time series of arthropod occurrences across gradients of land-use intensity with which to confirm causal relationships. Moreover, it remains unclear which land-use types and arthropod groups are affected, and whether the observed declines in biomass and diversity are linked to one another. Here we analyse data from more than 1 million individual arthropods (about 2,700 species), from standardized inventories taken between 2008 and 2017 at 150 grassland and 140 forest sites in 3 regions of Germany. Overall gamma diversity in grasslands and forests decreased over time, indicating loss of species across sites and regions. In annually sampled grasslands, biomass, abundance and number of species declined by 67%, 78% and 34%, respectively. The decline was consistent across trophic levels and mainly affected rare species; its magnitude was independent of local land-use intensity. However, sites embedded in landscapes with a higher cover of agricultural land showed a stronger temporal decline. In 30 forest sites with annual inventories, biomass and species number—but not abundance—decreased by 41% and 36%, respectively. This was supported by analyses of all forest sites sampled in three-year intervals. The decline affected rare and abundant species, and trends differed across trophic levels. Our results show that there are widespread declines in arthropod biomass, abundance and the number of species across trophic levels. Arthropod declines in forests demonstrate that loss is not restricted to open habitats. Our results suggest that major drivers of arthropod decline act at larger spatial scales, and are (at least for grasslands) associated with agriculture at the landscape level. This implies that policies need to address the landscape scale to mitigate the negative effects of land-use practices.

Much of the debate surrounding the human-induced biodiversity crisis has focused on vertebrates<sup>3</sup>, but population declines and extinctions may be even more substantial in small organisms such as terrestrial arthropods<sup>4</sup>. Recent studies have reported declines in the biomass of flying insects<sup>2</sup>, and in the diversity of insect pollinators<sup>5,6</sup>, butterflies and moths<sup>1,7–10</sup>, hemipterans<sup>11,12</sup> and beetles<sup>7,13,14</sup>. Owing to the associated negative effects on food webs<sup>15</sup>, ecosystem functioning and ecosystem services<sup>16</sup>, this insect loss has spurred an intense public debate. However, time-series data relating to arthropods are limited, and studies have so far focused on a small range of taxa<sup>11,13,14</sup>, a few types of land use and

habitat<sup>12</sup>—or even on single sites<sup>1,17</sup>. In addition, many studies lack species information<sup>2</sup> or high temporal resolution<sup>2,12</sup>. It therefore remains unclear whether reported declines in arthropods are a general phenomenon that is driven by similar mechanisms across land-use types, taxa and functional groups.

The reported declines are suspected to be caused mainly by human land use<sup>2</sup>. Locally, farming practices can affect arthropods directly by application of insecticides<sup>18,19</sup>, mowing<sup>20</sup> or soil disturbance, or indirectly via changes in plant communities through the application of herbicides or fertilizer<sup>21</sup>. Forestry practices can also affect local arthropod

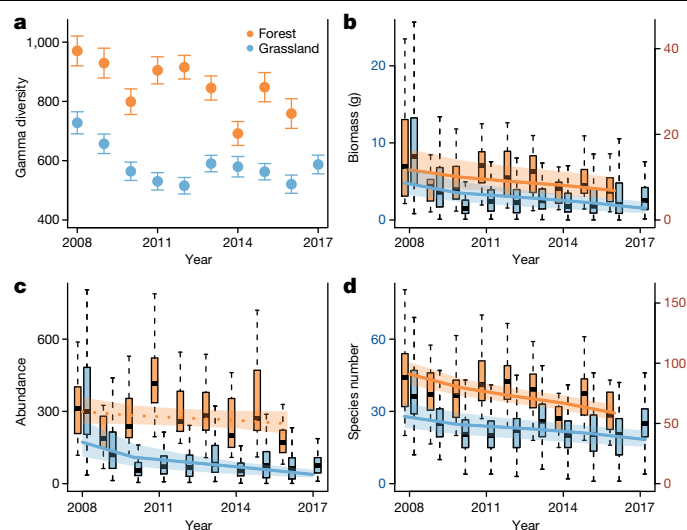
<sup>1</sup>Terrestrial Ecology Research Group, Department of Ecology and Ecosystem Management, Technical University of Munich, Freising, Germany. <sup>2</sup>Field Station Fabrikschleichach, Department of Animal Ecology and Tropical Biology, Julius-Maximilians-University Würzburg, Würzburg, Germany. <sup>3</sup>Forest Entomology, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland. <sup>4</sup>Ecological Networks, Department of Biology, Technical University of Darmstadt, Darmstadt, Germany. <sup>5</sup>Bavarian Forest National Park, Grafenau, Germany. <sup>6</sup>Department of Agricultural Biotechnology, Faculty of Agricultural and Natural Sciences, Düzce University, Düzce, Turkey. <sup>7</sup>Silviculture and Forest Ecology of the Temperate Zones, University of Göttingen, Göttingen, Germany. <sup>8</sup>Institute of Forest Sciences, Faculty of Environment and Natural Resources, University of Freiburg, Freiburg, Germany. <sup>9</sup>Institute of Plant Sciences, University of Bern, Bern, Switzerland. <sup>10</sup>Evolutionary Zoology, Biosciences, Salzburg University, Salzburg, Austria. <sup>11</sup>Department of Animal Ecology and Tropical Biology, Julius-Maximilians-University Würzburg, Würzburg, Germany. <sup>12</sup>Faculty of Geography, Philipps-University Marburg, Marburg, Germany. <sup>13</sup>Max Planck Institute for Biogeochemistry, Jena, Germany. \*e-mail: sebastian.seibold@tum.de

communities via changes in tree species composition or forest structure<sup>22</sup>. In addition, local arthropod populations can be affected by land use in the surrounding landscape; for example, through the drift and transport of pesticides and nitrogen by air or water<sup>23,24</sup>, through the effects of habitat loss on meta-communities (source–sink dynamics<sup>25</sup>) or by hampering dispersal.

To disentangle the local and landscape-level effects of land use on temporal trends in arthropod communities of grasslands and forests, we used data from the 'Biodiversity Exploratories' research programme that pertain to more than 1 million individual arthropods (2,675 species) (Extended Data Table 1). Arthropods were collected annually at 150 grassland sites by standardized sweep-net sampling in June and August from 2008 to 2017, and at 30 forest sites with flight-interception traps over the whole growing period from 2008 to 2016. An additional 110 forest sites were sampled in 2008, 2011 and 2014 to test for trends across a larger number of sites. Both the grassland and the forest sites cover gradients in local land-use intensity. Land-use intensity was quantified in the form of compound indices that are based on grazing, mowing and fertilization intensity in grasslands<sup>26</sup>, and on recent biomass removal, the proportion of non-natural tree species and deadwood origin in forests<sup>27</sup>. To analyse landscape-level effects, we quantified the cover of arable fields, grassland and forest in circles, with a radius between 250 m and 2 km, around each sampling site. We modelled temporal trends in arthropod biomass (estimated from body size; Methods), abundance and the number of species separately for grasslands and forests, and tested for the effects of local and landscape-scale land-use intensity on these trends, accounting for weather conditions. Analyses were conducted for all species together, and for different dispersal and trophic guilds.

The total number of arthropod species across all sites (gamma diversity) was substantially lower in later than in earlier years in both forests and grasslands (Fig. 1). Gamma diversity, biomass, abundance and number of species fluctuated over time but revealed an overall decrease with strongest declines from 2008 to 2010, especially in grasslands (Fig. 1). Year-to-year fluctuations in arthropod biomass, abundance and number of species were partially explained by weather conditions (Extended Data Fig. 1, Supplementary Table 1-1, Supplementary Information section 2). Accounting for weather, fitted trends from our models showed declines in biomass of 67% for grasslands and 41% for forests, declines in species numbers of 34% for grasslands and 36% for forests, and declines in abundance of 78% for grasslands, with no significant change in abundances for forests (−17%) (Fig. 1, Supplementary Table 3-1). In grasslands, declines occurred consistently across all trophic guilds (herbivores, myceto-detritivores, omnivores and carnivores), although the trend for carnivores was not significant (Supplementary Table 1-1). In forests, the patterns were more complex: herbivores showed an increase in abundance and species number, whereas all other trophic guilds declined. Temporal trends of arthropods on the basis of data recorded in 3-year intervals from all 140 forest sites were similar to the trends based on the 30 sites with annual data (Supplementary Table 1-1). Sensitivity analyses that removed or reshuffled years showed that the decline was influenced by, but not solely dependent on, high numbers of arthropods in 2008. Fluctuations in numbers (including the numbers from 2008) appear to match trends that have been observed in other studies<sup>2</sup>, which suggests that the recent decline is part of a longer-term trend that had begun by at least the early 1990s (Extended Data Fig. 2, Supplementary Information section 3). Further sensitivity analyses showed consistent declines when data from individual sampling dates were not aggregated per year, and also showed that declines concerned all three regions that we analysed (Supplementary Tables 3-2, 3-3, Supplementary Fig. 3-1).

Linking changes in biomass, abundance and the number of species to one another enables further inferences regarding the mechanisms that drive arthropod declines. In grasslands, both abundant and less-abundant species declined in abundance (Fig. 2), but loss in the number of species occurred mostly among less-frequent species (Fig. 1, Extended Data Fig. 3, Supplementary Information section 4). This suggests that the



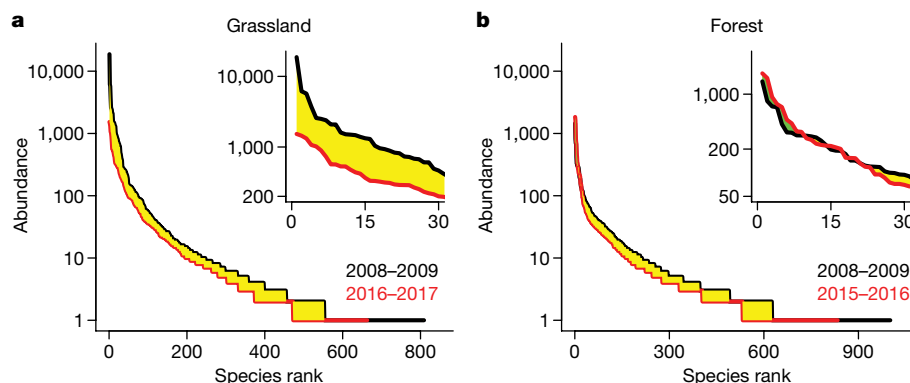
**Fig. 1 | Temporal trends in arthropod communities.** **a–d**, Gamma diversity (total number of species across all grassland or forest sites) (**a**), biomass (**b**), abundance (**c**) and number of species (**d**) of arthropods were recorded in 30 forest and 150 grassland sites across Germany. Gamma diversity shows mean incidence-based, bias-corrected diversity estimates (Chao's BSS, that is, the higher value of the minimum doubled reference sample size and the maximum reference sample size among years<sup>29</sup>) for  $q = 0$  and 95% confidence intervals derived from bootstrapping ( $n = 200$ ). Non-overlapping confidence intervals indicate significant difference<sup>30</sup>. Box plots show raw data per site and year ( $n = 1,406$  (grassland) or 266 (forest) independent samples). Solid lines indicate significant temporal trends ( $P < 0.05$ ) based on linear mixed models that included weather conditions, and local and landscape-level land-use intensity as covariates. Shaded areas represent confidence intervals. Boxes represent data within the 25th and 75th percentile, black lines show medians, and whiskers show 1.5× the interquartile range. Data points beyond that range (outliers) are not shown for graphical reasons. Plots for biomass and species number have separate y axes for grassland and forest.

decline in the number of species in grasslands was attributable mainly to a loss of individuals among rare species. In forests, species that were initially less abundant decreased in abundance, whereas some of the most abundant species—including invasive species and potential pest species—increased in abundance (Fig. 2, Supplementary Table 5-1). The loss of species was, however, irrespective of their frequency (Fig. 1, Extended Data Fig. 3, Supplementary Information section 4). This suggests that the decline of arthropods in forests is driven by mechanisms that negatively affect the abundances of many species, which leads to an overall decline in biomass and the number of species but favours some species that are able to compensate declines in abundance.

The magnitudes of declines in biomass, abundance and the number of species in arthropod communities were independent of local land-use intensity (Supplementary Table 1-1) as well as changes in plant communities (Supplementary Information section 6) at all sites. However, in forests declines in the number of species were weaker at sites with high natural or anthropogenic tree mortality, possibly owing to increased heterogeneity in local habitats (Extended Data Fig. 4). Landscape composition had no effect on arthropod trends in forests (note that forest sites covered only limited gradients of the landscape variables, Extended Data Fig. 5), but it mediated declines in the number of species in grasslands: the magnitude of the declines increased with increasing cover of arable fields, and marginally increased with cover of grasslands in the surrounding landscape (Fig. 3, Supplementary Table 1-1). This suggests that major drivers of arthropod decline in grasslands are associated with agricultural land use at the landscape scale.

The interaction between a species and the landscape around its habitat depends on its dispersal ability, which ultimately determines



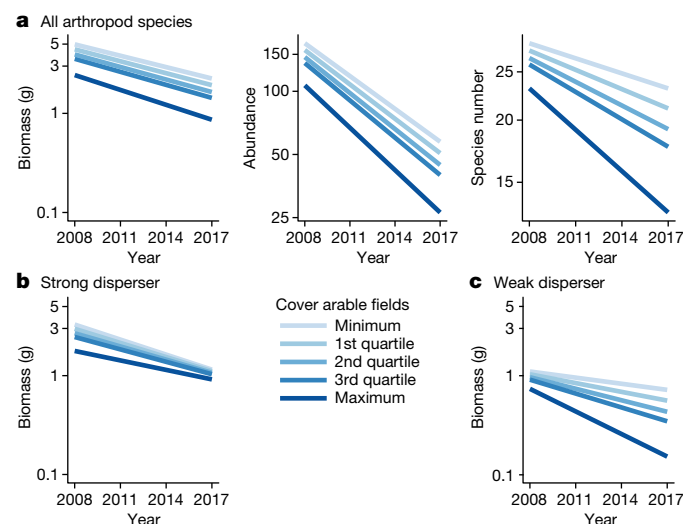


**Fig. 2 | Changes in the dominance of species.** Rank abundance curves of arthropod communities for the first two (2008–2009) and final two (2016–2017 for grasslands and 2015–2016 for forests) years of the study, from 150 grassland and 30 forest sites. The insets show enlarged curves for the 30 most abundant species. Data from the first two and final two study years were pooled

(abundances are the total number of individuals of a species observed over two years). Declines in abundance are highlighted by yellow shading, and increases in abundance are shaded in green. The y axes are log-scaled, but show untransformed values.

its occurrence and persistence<sup>28</sup>. In grasslands, taxa of high and low dispersal ability (Methods) both declined, but an increasing cover of arable fields—although not of grasslands—in the surroundings amplified declines in the biomass of weak dispersers more strongly than it did declines of strong dispersers (Fig. 3, Supplementary Table 7-1). Weak dispersers may experience higher mortality during dispersal, and thus have a lower chance of (re)colonization of a particular site when arable field cover is high. In forests, strong dispersers declined in biomass, abundance and the number of species, whereas weak dispersers increased in abundance and biomass—but less strongly when grassland cover in the landscape was high (Supplementary Table 7-1). This suggests that the drivers behind arthropod declines in forests also act at landscape-level spatial scales.

We showed that arthropods declined markedly not only in biomass but also in abundance and the number of species, and that this affected taxa of most trophic levels in both grasslands and forests. Declines in gamma diversity suggest that species might disappear across regions.



**Fig. 3 | Landscape effects on arthropod decline in grasslands.** **a**, Temporal changes in biomass, abundance and the number of species for all arthropod species. **b, c**, Temporal change in biomass of species with high (**b**) or low (**c**) dispersal ability, conditional on the cover of arable fields in the surrounding landscape (1-km radius). The decline in biomass increased significantly with the cover of arable fields for weak dispersers, but not for strong dispersers. Slopes were derived from models that included weather conditions and local land-use intensity as covariates. The y axes are log-scaled, but show untransformed values.

Our results also indicate that the major drivers of arthropod decline in both habitat types act at landscape-level spatial scales, but that declines may be moderated by increases in heterogeneity of local habitats in forests. Although the drivers of arthropod decline in forests remain unclear, in grasslands these drivers are associated with the proportion of agricultural land in the landscape. However, we cannot ascertain whether the observed declines are driven by the legacy effects of historical land-use intensification or by recent agricultural intensification at the landscape level; for example, by the decrease of fallow land and field margins rich in plant species, the increased use of pesticides or use of more potent insecticides (Supplementary Information section 3). Time-series data relating to changes in the use of agrochemicals or the presence of fine-scale arthropod habitats would be necessary to answer this question. Furthermore, the extents to which changes in climate have reinforced the observed trends in arthropod biomass, abundance and number of species is unclear (Supplementary Information section 2). Our results show that widespread arthropod declines have occurred in recent years. Although declines were less pronounced during the second half of our study period, there is no indication that negative trends have been reversed by measures that have been implemented in recent years. This calls for a paradigm shift in land-use policy at national and international levels to counteract species decline in open and forested habitats by implementing measures that are coordinated across landscapes and regions. Such strategies should aim to improve habitat quality for arthropods and to mitigate the negative effects of land-use practices not only at a local scale (within isolated patches embedded in an inhospitable agricultural matrix) but also across large and continuous areas.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1684-3>.

1. Habel, J. C. et al. Butterfly community shifts over two centuries. *Conserv. Biol.* **30**, 754–762 (2016).
2. Hallmann, C. A. et al. More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PLoS ONE* **12**, e0185809 (2017).
3. Ceballos, G. et al. Accelerated modern human-induced species losses: entering the sixth mass extinction. *Sci. Adv.* **1**, e1400253 (2015).
4. Thomas, J. A. et al. Comparative losses of British butterflies, birds, and plants and the global extinction crisis. *Science* **303**, 1879–1881 (2004).
5. Biesmeijer, J. C. et al. Parallel declines in pollinators and insect-pollinated plants in Britain and the Netherlands. *Science* **313**, 351–354 (2006).

6. Potts, S. G. et al. Global pollinator declines: trends, impacts and drivers. *Trends Ecol. Evol.* **25**, 345–353 (2010).
7. Dirzo, R. et al. Defaunation in the Anthropocene. *Science* **345**, 401–406 (2014).
8. Conrad, K. F., Warren, M. S., Fox, R., Parsons, M. S. & Woiwod, I. P. Rapid declines of common, widespread British moths provide evidence of an insect biodiversity crisis. *Biol. Conserv.* **132**, 279–291 (2006).
9. Maes, D. & Van Dyck, H. Butterfly diversity loss in Flanders (north Belgium): Europe's worst case scenario? *Biol. Conserv.* **99**, 263–276 (2001).
10. Stefanescu, C., Torre, I., Jubany, J. & Páramo, F. Recent trends in butterfly populations from north-east Spain and Andorra in the light of habitat and climate change. *J. Insect Conserv.* **15**, 83–93 (2011).
11. Schuch, S., Wesche, K. & Schaefer, M. Long-term decline in the abundance of leafhoppers and planthoppers (Auchenorrhyncha) in Central European protected dry grasslands. *Biol. Conserv.* **149**, 75–83 (2012).
12. Schuch, S., Bock, J., Krause, B., Wesche, K. & Schaefer, M. Long-term population trends in three grassland insect groups: a comparative analysis of 1951 and 2009. *J. Appl. Entomol.* **136**, 321–331 (2012).
13. Brooks, D. R. et al. Large carabid beetle declines in a United Kingdom monitoring network increases evidence for a widespread loss in insect biodiversity. *J. Appl. Ecol.* **49**, 1009–1019 (2012).
14. Desender, K., Dekoninck, W., Dufrière, M. & Maes, D. Changes in the distribution of carabid beetles in Belgium revisited: have we halted the diversity loss? *Biol. Conserv.* **143**, 1549–1557 (2010).
15. Bowler, D. E., Heldbjerg, H., Fox, A. D., de Jong, M. & Böhning-Gaese, K. Long-term declines of European insectivorous bird populations and potential causes. *Conserv. Biol.* **33**, 1120–1130 (2019).
16. Siemann, E. & Weisser, W. W. (eds) *Insects and Ecosystem Function* (Springer, 2004).
17. Shortall, R. C. et al. Long-term changes in the abundance of flying insects. *Insect Conserv. Divers.* **2**, 251–260 (2009).
18. Geiger, F. et al. Persistent negative effects of pesticides on biodiversity and biological control potential on European farmland. *Basic Appl. Ecol.* **11**, 97–105 (2010).
19. Ewald, J. A. et al. Influences of extreme weather, climate and pesticide use on invertebrates in cereal fields over 42 years. *Glob. Chang. Biol.* **21**, 3931–3950 (2015).
20. Gossner, M. M. et al. Land-use intensification causes multitrophic homogenization of grassland communities. *Nature* **540**, 266–269 (2016).
21. Haddad, N. M., Haarstad, J. & Tilman, D. The effects of long-term nitrogen loading on grassland insect communities. *Oecologia* **124**, 73–84 (2000).
22. Penone, C. et al. Specialisation and diversity of multiple trophic groups are promoted by different forest features. *Ecol. Lett.* **22**, 170–180 (2019).
23. Brittain, C. A., Vighi, M., Bommarco, R., Settele, J. & Potts, S. G. Impacts of a pesticide on pollinator species richness at different spatial scales. *Basic Appl. Ecol.* **11**, 106–115 (2010).
24. de Jong, F. M. W., de Snoo, G. R. & van de Zande, J. C. Estimated nationwide effects of pesticide spray drift on terrestrial habitats in the Netherlands. *J. Environ. Manage.* **86**, 721–730 (2008).
25. Thomas, J. A. et al. The quality and isolation of habitat patches both determine where butterflies persist in fragmented landscapes. *Proc. R. Soc. Lond. B* **268**, 1791–1796 (2001).
26. Blüthgen, N. et al. A quantitative index of land-use intensity in grasslands: Integrating mowing, grazing and fertilization. *Basic Appl. Ecol.* **13**, 207–220 (2012).
27. Kahl, T. & Bauhus, J. An index of forest management intensity based on assessment of harvested tree volume, tree species composition and dead wood origin. *Nat. Conserv.* **7**, 15–27 (2014).
28. Tscharntke, T. et al. Landscape moderation of biodiversity patterns and processes – eight hypotheses. *Biol. Rev. Camb. Philos. Soc.* **87**, 661–685 (2012).
29. Chao, A. et al. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol. Monogr.* **84**, 45–67 (2014).
30. Schenker, N. & Gentleman, J. F. On judging the significance of differences by examining the overlap between confidence intervals. *Am. Stat.* **55**, 182–186 (2001).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

### Study system, land-use measures and weather data

The study was conducted as part of the Biodiversity Exploratories project ([www.biodiversity-exploratories.de](http://www.biodiversity-exploratories.de)) in three regions of Germany: (1) Schwäbische Alb in southwestern Germany (460–860 m above sea level (asl)); (2) Hainich-Dün in central Germany (285–550 m asl); and (3) Schorfheide-Chorin in northeastern Germany (3–140 m asl). The three regions differ in climate, geology and topography, but each is characterized by gradients of land-use intensity in grasslands and forests that are typical for large parts of temperate Europe<sup>31</sup>.

A total of 150 grassland sites of 50 × 50 m in size (50 per region) and 140 forest sites of 100 × 100 m in size (49 in Schwäbische Alb, 50 in Hainich-Dün and 41 in Schorfheide-Chorin), located within larger management units, were selected from a total of about 3,000 candidate sites by stratified random sampling to ensure that the selected sites covered the whole range of land-use intensity and to minimize the confounding effects of spatial position or soil type<sup>31</sup>. All sites have a long history of the same type of land-use and of broadly similar land-use intensity. Land use is conducted by landowners or tenant farmers (rather than by the scientific consortium) and the start of the project did not cause changes in land use. Local land-use intensity decreased significantly in forests and marginally in grasslands over the course of our study (Supplementary Information section 8). No pesticides were applied at any of the sites, except for application of herbicides in grasslands in five occasions (site number and year: AEG2 2011, HEG2 2013, HEG36 2014, HEG37 2014 and HEG1 2015).

In grasslands, the gradient of land-use intensity ranged from semi-natural to intensively managed grasslands. Natural grasslands, which do not require management to prevent succession to forest, are almost entirely absent from western and central Europe. All sites were continuously managed by farmers. Information on management practices, including the level of fertilization ( $\text{kg N ha}^{-1} \text{ year}^{-1}$ ), grazing (number of livestock units  $\text{ha}^{-1} \text{ year}^{-1}$ ) and mowing (number of cuts  $\text{year}^{-1}$ ), was assessed annually by standardized interviews with the farmers. Local land-use intensity in grasslands was then quantified as a compound index by summing the standardized intensities (that is, divided by the global mean value) of these three components<sup>26</sup>. We then calculated the mean local land-use intensity for each site over the ten years of our study (2008–2017). The least intensively managed grasslands are often located within protected areas ( $n = 47$  sites, including 15 sites in strictly protected areas) and are typically grazed by 40–50 sheep per hectare for about 10 days per year (or more rarely by 1–3 cattle per hectare for 20 days), unfertilized and not mown. Grasslands of intermediate land-use intensity are usually unfertilized (or fertilized with less than 30  $\text{kg N ha}^{-1} \text{ year}^{-1}$ ), and are either mown twice a year or grazed by 4 cattle per hectare for about 50 days. The most intensively managed grasslands are typically fertilized (60–120  $\text{kg N ha}^{-1} \text{ year}^{-1}$ ) and either mown 2 or 3 (maximum of 4) times a year, grazed by 5–10 cattle per hectare for 100–150 days, or both mown and grazed.

In forests, the gradient of land-use intensity included three broad categories: unmanaged broadleaf, managed broadleaf and managed conifer forest. The least intensively managed forests have been managed to some degree in the past, but are now often located within protected areas ( $n = 56$  sites, including 31 sites in strictly protected areas; 14 and 9, respectively, of these sites have annual arthropod data). The naturally dominant tree species at all sites is European beech (*Fagus sylvatica*). The dominating conifer species are Norway spruce (*Picea abies*) and Scots pine (*Pinus sylvestris*), which are native to central Europe but would be absent or rare in the study regions under natural conditions. However, these trees have been cultivated in the study regions for approximately

250 years. On all sites, inventories of living trees, stumps and deadwood were conducted between 2009 and 2011<sup>27</sup>. To obtain a continuous fine-grained measure for local land-use intensity in forests, we calculated a compound index on the basis of three indicators scaled 0–1: recent biomass removal (volume of harvested timber divided by the sum of the volume of living trees, harvested timber and remaining deadwood), proportion of tree species that do not belong to the native vegetation community (volume of standing timber, harvested timber and deadwood of non-native tree species (including spruce and pine) divided by the sum of the volume of all tree species) and deadwood origin (volume of deadwood with saw cuts divided by the total volume of deadwood)<sup>27</sup>.

Land-use intensity at the landscape scale was quantified by measuring the proportion of area covered by arable fields, grasslands and forests within circular areas around the centre of our sites. As the scale of effect was unknown, we considered different area sizes with radii 250, 500, 1,000, 1,500 and 2,000 m. We used vector land-cover data from ATKIS Basis DLM (license agreements: GeoBasis-DE/LGB 2017, BG-D 29/17) with  $\pm 3$  m accuracy of polygon borders, representing conditions between 2008 and 2010. We are aware that land cover is only a coarse measure for land-use intensity at the landscape scale, but information on fine-scaled habitat availability (such as field margins and within-field plant diversity) or details on land-use practices (such as the amount of fertilizers and pesticides used) was not available.

Air temperature was recorded at all 290 sites with hourly resolution starting between early 2008 and early 2009, depending on the site. Gaps within the time series at individual stations were filled on the basis of average linear relationships with neighbouring stations within the three regions. To derive complete time series from winter 2007–2008 onward, the initial time span was filled on the basis of data from the station network of the German Weather Service that surrounds each Exploratory (five stations each). Using 10 × space–time cross-validation and a forward feature selection approach, the best individual subset of the 5 surrounding stations for each of the 290 stations to be filled was identified and a multiple linear model was used to predict the missing values. Precipitation for site was derived from the RADOLAN product of the German Weather Service (hourly radar-based precipitation estimates corrected by gauge measurements, with a resolution of 1  $\text{km}^2$  and 0.1  $\text{mm/h}$ ). From these products, we calculated—for each region and year—the mean temperature, number of frost days (daily minimum temperature  $< 0^\circ\text{C}$ ), number of warm days (daily mean temperature  $> 20^\circ\text{C}$ ) and precipitation sum for winter (from November of the previous year to February), growing period (from March to October) and year (from November of the previous year to October). Gap filling at the start of the time series was conducted in R version 3.5.1<sup>32</sup>. Other computations used the climate-processing software TubeDB (<https://environmentalinformatics-marburg.github.io/tubedb>).

### Arthropod sampling

Arthropods were sampled annually in a consistent and standardized way from all 150 grassland sites from 2008 to 2017 and from 30 forest sites from 2008 to 2016. On the other 110 forest sites, arthropods were sampled by the same method and with the same sampling effort in 2008, 2011 and 2014. In grasslands, all arthropods of the herb layer were sampled twice per year in June and August to represent different phenological windows within the peak season of adult arthropod activity. On the basis of monthly samplings at the beginning of the study, we identified these two months as representing the best trade-off between reducing sampling effort and covering most species. Arthropods were sampled by sweep netting along a 150-m-long transect that comprised 3 of the virtual borders of a site by conducting 60 double sweeps per site<sup>20</sup>. Sweep netting was conducted only on days without rain, with low wind speed and after morning dew had dried. To reduce potential observer bias, personnel were trained and changes in personnel were reduced as much as possible (one change in Schwäbische Alb and Hainich-Dün regions in 2009; two changes in Schorfheide-Chorin region in 2009 and

2010; Supplementary Information section 3). In forests, flying insects were sampled using two flight-interception traps per site located close to two randomly selected corners of each site. Traps consisted of two crossed transparent plastic shields (40 × 60 cm) with funnels opening into sampling jars below and above the shields that were filled with 3% CuSO<sub>4</sub> solution and a drop of detergent<sup>33</sup>. Traps operated from March to October and were emptied monthly. All sites are part of larger management units, and therefore no edge effects owing to changes in land-use intensity at site borders are to be expected.

All samples were sorted to the order level in the laboratory. For taxonomic groups that occurred in larger numbers, and for which expert taxonomists were available, adult specimens were identified at species level: for grasslands, these included species in the Araneae, Coleoptera, Hemiptera (both Heteroptera and Auchenorrhyncha; some hemipterans were classified only to family or subfamily level) and Orthoptera, and for forests, these included species in the Coleoptera and Hemiptera (in Heteroptera). Only very few adults in these taxonomic groups could not be identified to the species level (1.1% in grasslands and 0.7% in forests), and these were excluded from the analyses. In grasslands, we additionally counted the number of individuals per order for groups for which no taxonomists were available: these included Acarina, Blattodea, Collembola, Dermaptera, Diptera (which was divided into Brachycera and Nematocera), Ephemeroptera, Hemiptera (Psyllidae and Aphidoidea), Hymenoptera (divided into Apocrita, Symphyta and Formicidae), Isopoda, Lepidoptera, Mecoptera, Megaloptera, Neuroptera, Odonata, Opiliones, Plecoptera, Pseudoscorpiones, Psocoptera, Raphidoptera, Strepsiptera, Thysanoptera and Trichoptera. Information on body length, trophic level and dispersal ability for identified species was obtained from the literature<sup>34,35</sup>. We estimated the biomass of all arthropod specimens that were identified to species level by applying a previously developed<sup>36</sup> general power function: biomass (in g) =  $0.305 \times L^{2.62} / 1,000$  where  $L$  is the mean body length of a species in millimetres. All arthropods identified to the species level were assigned to one of four trophic groups (herbivores, myceto-detritivores, carnivores and omnivores) on the basis of their known main food resource as adults. Because typical dispersal distances are unknown for most arthropod species, we classified species according to morphological characteristics and behavioural traits within taxonomic groups (for example, wing development, ballooning or hunting strategy)<sup>35</sup>. Dispersal ability—ranging from 0 to 1 in steps of 0.25—was defined differently for the groups, considering wing dimorphism (for Hemiptera, Coleoptera and Orthoptera), flying ability (Coleoptera) as well as information on migration and ballooning behaviour (Araneae) (details have previously been published<sup>34</sup>). All species with a dispersal ability ≤ 0.5 were considered to be weak dispersers, and all species with a dispersal ability > 0.5 were considered to be strong dispersers.

## Vegetation sampling

Plant communities in all 150 grassland sites were recorded in an area of 4 × 4 m between mid-May and mid-June from 2008 to 2017, and in 30 forest sites in an area of 20 × 20 m twice a year (spring and summer) from 2009 to 2016, by estimating the cover of each species. Ellenberg indicator values were taken from a previous publication<sup>37</sup>.

## Statistical analyses

All analyses were conducted in R v.3.5.1<sup>32</sup>.

We performed principal component analyses (PCAs) and pairwise correlation tests including all weather variables. On the basis of the results, and similar to a previous publication<sup>2</sup>, we selected mean winter temperature and precipitation during the growing period for subsequent analyses because these variables were weakly correlated and represented both seasons and both temperature and precipitation (Extended Data Fig. 6). Despite considerable temporal fluctuations, mean winter temperature increased and precipitation during the growing period decreased within our study period (Extended Data Fig. 7). PCAs and pairwise correlation tests for landscape-level variables enabled

us to select cover of arable fields and cover of grassland as independent predictors because these were weakly correlated across spatial scales, whereas forest cover was correlated to both cover of arable fields and grassland (Extended Data Fig. 6).

We calculated gamma diversity (estimated total number of species) across 150 grassland and 30 forest sites separately for grassland and forest for each year using the ‘diversity accumulation curve’ framework that extends methods for rarefaction and extrapolation of species richness<sup>29</sup>. We used Chao’s BSS based on the frequencies of species: the higher value of the minimum doubled reference sample size, and the maximum reference sample size among years as incidence-based, bias-corrected diversity estimates for  $q = 0$ , 1 and 2<sup>29</sup>. This approach accounts for slight differences in site numbers between years caused by limited accessibility or failure of traps. With increasing order  $q$ , the more frequent species are more strongly weighted ( $q = 0$  equals species richness,  $q = 1$  equals the exponential of Shannon entropy and  $q = 2$  equals the inverse of Simpson diversity), which enables us to assess whether changes in gamma diversity depend on the frequencies of species. Using different reference-sample sizes resulted in consistent results (data not shown). Confidence intervals were calculated by bootstrapping ( $n = 200$  bootstraps).

We aggregated data from all arthropods identified to species per site and year to calculate biomass, abundance and the number of species for all species, and separately for each trophic and both of the dispersal groups. For grasslands, we additionally calculated the abundance of all arthropods per site and year, including groups that were not identified to species level. To identify the scale of the effect for landscape-level land-use intensity<sup>38</sup>, we conducted a multiscale analysis by correlating arthropod biomass, abundance and the number of species with the cover of arable fields and cover of grassland separately for radii of 250–2,000 m. For this, only data from a random subset of sites with non-overlapping buffers at the 2,000-m scale were used, and this procedure was repeated 100 times. In grasslands, correlations increased initially with increasing radius but started to plateau at 1,000 m (Extended Data Fig. 8). Owing to the higher overlap of buffers of neighbouring sites at larger spatial scales, we thus present results for all grassland analyses at the 1,000-m scale. In forests, the patterns were more complex, but—because of the small range of agricultural land-use variables at small scales (Extended Data Fig. 5) and the higher overlap of buffers of neighbouring sites at larger spatial scales—we also present the results for all forest analyses at the 1,000-m scale.

To test for temporal trends in our arthropod data, we fitted generalized linear mixed models with Poisson errors for count data (abundance and species number; function `glmer` in package `lme4`) and linear mixed models with Gaussian errors for biomass (log-transformed; function `lmer`), separately for grasslands and forests. For forests, we analysed the annual data from 30 sites and the 3-year-interval data from 140 sites separately. Separate models were fitted for trophic groups. Fixed effects included year, weather (mean winter temperature, precipitation during the growing period and their interaction), local land-use intensity and landscape-level land-use intensity (cover of arable fields and cover of grassland within a radius of 1,000 m), as well as interactions between year and local land-use intensity and between year and landscape-level land-use intensities. Models included the site nested in the region as a random effect to account for the nested design and the repeated measures at the site level. Poisson models included an observation-specific random effect to account for potential overdispersion<sup>39</sup>. All continuous predictor variables were standardized to a mean of 0 and an s.d. of 1 before modelling. To test whether changes in the overall number of species were associated with changes in overall abundance, we ran additional models with the number of species as response and log-transformed abundance as covariate. To assess the contribution of individual years to the overall trend, we repeated the models for overall biomass, abundance and number of species, and excluded data from one year each time. In addition, we tested whether the observed effect



of year differed from a random expectation by randomizing the order of years 100× for forests and grasslands before modelling.

To test for differences between dispersal groups, we fitted models for biomass, abundance and number of species in which effects of year, local and landscape-level land-use intensity (as well as their interactions) were estimated specifically for each dispersal guild. These models included response values for each group per site and year, and dispersal group (weak or strong) as fixed effect. To test whether observed effects differed significantly between dispersal guilds, we fitted additional models including the three three-way interactions between dispersal guild, year and each of the three land-use variables. All models included the site nested in the region as a random effect to account for spatial arrangement and temporal repetitions per site. Poisson models included an observation-specific random effect to account for potential overdispersion.

In addition to models for data aggregated per site and year, we fitted models for biomass, abundance and number of species at the level of individual observations (two collections per year for grasslands and five collections per year for forests), which could account for seasonal differences and weather conditions at the time of sampling. For forest data from 30 sites, fixed effects included mean winter temperature, mean temperature and precipitation during sampling period, length of sampling period (in days), Julian date of the day on which traps were emptied, local and landscape-level land-use intensity (cover of arable fields and cover of grassland within a radius of 1,000 m), as well as interactions between year and local land-use intensity, and between year and landscape-level land-use intensity. For grasslands, fixed effects included mean winter temperature, precipitation during the growing season and their interaction, mean temperature and precipitation on the day of sampling, Julian date of the day of sampling, local land-use intensity and landscape-level land-use intensity (cover of arable fields and cover of grassland within a radius of 1,000 m), as well as interactions between year and local land-use intensity, and between year and landscape-level land-use intensity. Models included the site nested in the region as a random effect to account for the nested design and the repeated measures at the site level. Poisson models included an observation-specific random effect to account for potential overdispersion<sup>39</sup>. To allow nonlinear effects for day of sampling, we fitted generalized additive models (function `gamm4` in package `gamm4`).

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

This work is based on data from several projects of the Biodiversity Exploratories programme (DFG Priority Program 1374). All data used for analyses are publicly available from the Biodiversity Exploratories

Information System (<https://doi.org/10.17616/R32P9Q>) at <https://www.bexis.uni-jena.de/PublicData/PublicDataSet.aspx?DatasetId=25786>. Raw data are publicly available from the same repository (with identifiers 21969, 22007, 22008, 19686 and 20366), or will become publicly available after an embargo period of five years from the end of data assembly to give the owners and collectors of the data time to perform their analysis. Any other relevant data are available from the corresponding author upon reasonable request.

31. Fischer, M. et al. Implementing large-scale and long-term functional biodiversity research: the Biodiversity Exploratories. *Basic Appl. Ecol.* **11**, 473–485 (2010).
32. R Core Team. *R: A Language and Environment for Statistical Computing* (2018).
33. Schall, P. et al. The impact of even-aged and uneven-aged forest management on regional biodiversity of multiple taxa in European beech forests. *J. Appl. Ecol.* **55**, 267–278 (2018).
34. Gossner, M. M. et al. A summary of eight traits of Coleoptera, Hemiptera, Orthoptera and Araneae, occurring in grasslands in Germany. *Sci. Data* **2**, 150013 (2015).
35. Birkhofer, K. et al. Land-use type and intensity differentially filter traits in above- and below-ground arthropod communities. *J. Anim. Ecol.* **86**, 511–520 (2017).
36. Rogers, L. E., Hinds, W. T. & Buschbom, R. L. A General weight vs. length relationship for insects. *Ann. Entomol. Soc. Am.* **69**, 387–389 (1976).
37. Ellenberg, H. et al. Zeigerwerte von Pflanzen in Mitteleuropa. *Scr. Geobot.* **18**, 248 (1991).
38. Fahrig, L. Rethinking patch size and isolation effects: the habitat amount hypothesis. *J. Biogeogr.* **40**, 1649–1663 (2013).
39. Elston, D. A., Moss, R., Boulinier, T., Arrowsmith, C. & Lambin, X. Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology* **122**, 563–569 (2001).
40. Hilmers, T. et al. Biodiversity along temperate forest succession. *J. Appl. Ecol.* **55**, 2756–2766 (2018).

**Acknowledgements** We thank T. Lewinsohn, S. Meyer and V. Wolters for their comments and suggestions for the analyses; M. Lutz, J. Bartzeko, P. Freynhagen, I. Gallenberger, M. Türke, M. Lange, T. Kahl, E. Pašalić, E. Sperr, K. Kremer and all student helpers for conducting arthropod sampling in the field and laboratory; R. Achziger, E. Anton, T. Blick, B. Büche, M.-A. Fritze, R. Heckmann, A. Kästner, F. Köhler, G. Köhler, T. Köllebeck, C. Morkel, F. Schmolke, T. Wagner and O. Wiche for arthropod species identification; C. Seilwinder and R. Honecker for GIS work; the managers of the three Exploratories (K. Wells, S. Renner, K. Reichel-Jung, S. Gockel, K. Wiesner, K. Lorenzen, A. Hemp and M. Gorke) for their work in maintaining the site and project infrastructure; C. Fischer and S. Pfeiffer for giving support through the central office; A. Ostrowski, M. Owonibi and J. Nieschulze for managing the central database; and D. Hessenmöller, I. Schöning, F. Buscot and the late E. Kalko for their role in setting up the Biodiversity Exploratories project. The work has been funded by the DFG Priority Program 1374 'Infrastructure-Biodiversity-Exploratories'. Field work permits were issued by the responsible state environmental offices of Baden-Württemberg, Thüringen and Brandenburg.

**Author contributions** S.S., J.M. and N.K.S. conceived the idea for the manuscript; M.M.G., N.K.S., S.S., D.A., W.W.W., T.N., S.W., P.S., C.A., J.B., J.V., D.P. and M.F. collected and processed data; S.S., J.M., M.M.G. and W.W.W. defined the final analysis; S.S., N.K.S., C.P., P.S. and M.M.G. analysed the data; S.S. and W.W.W. wrote the first manuscript draft and finalized the manuscript. All authors discussed the analyses and commented on the manuscript.

**Competing interests** The authors declare no competing interests.

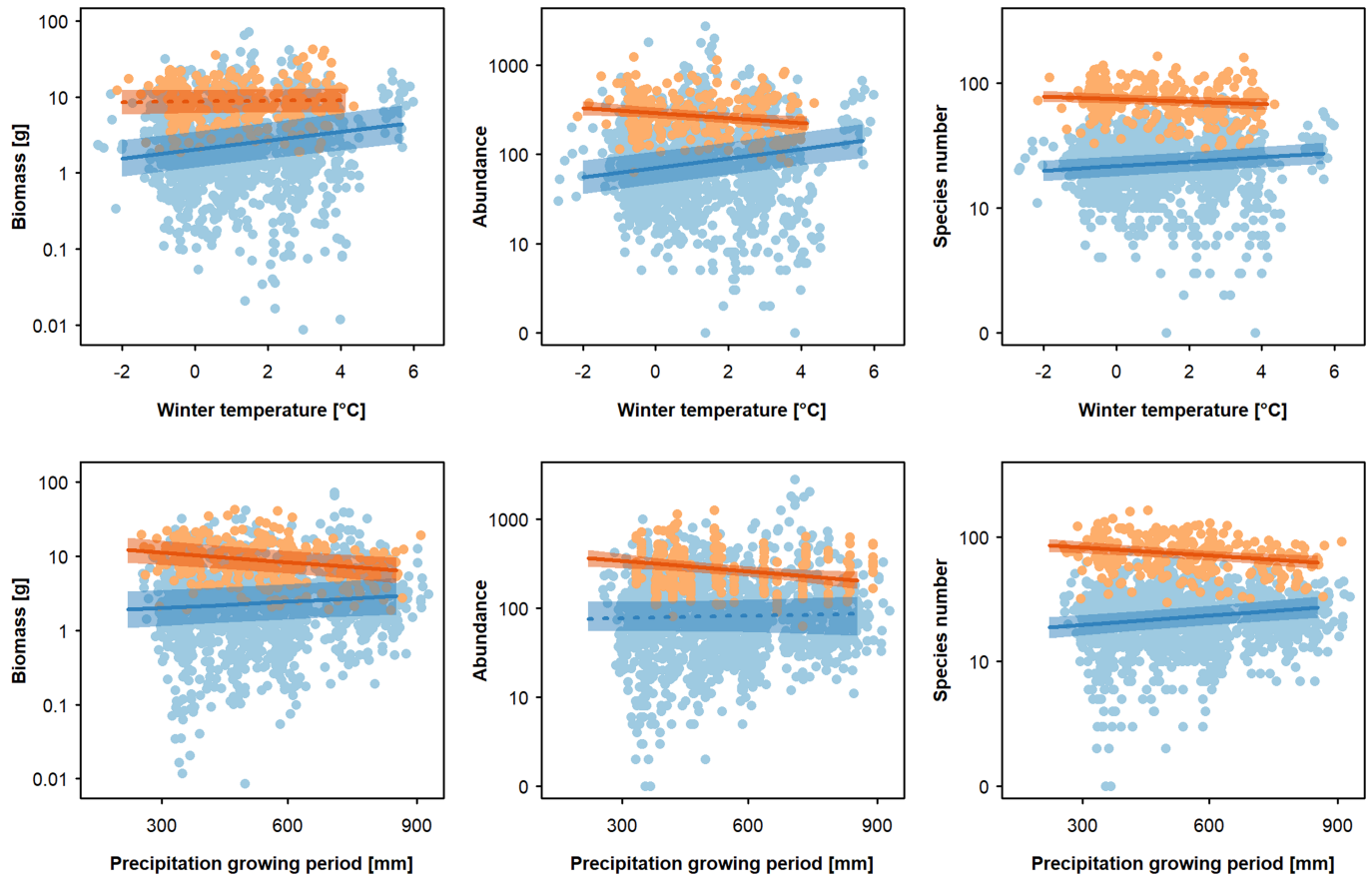
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1684-3>.

**Correspondence and requests for materials** should be addressed to S.S.

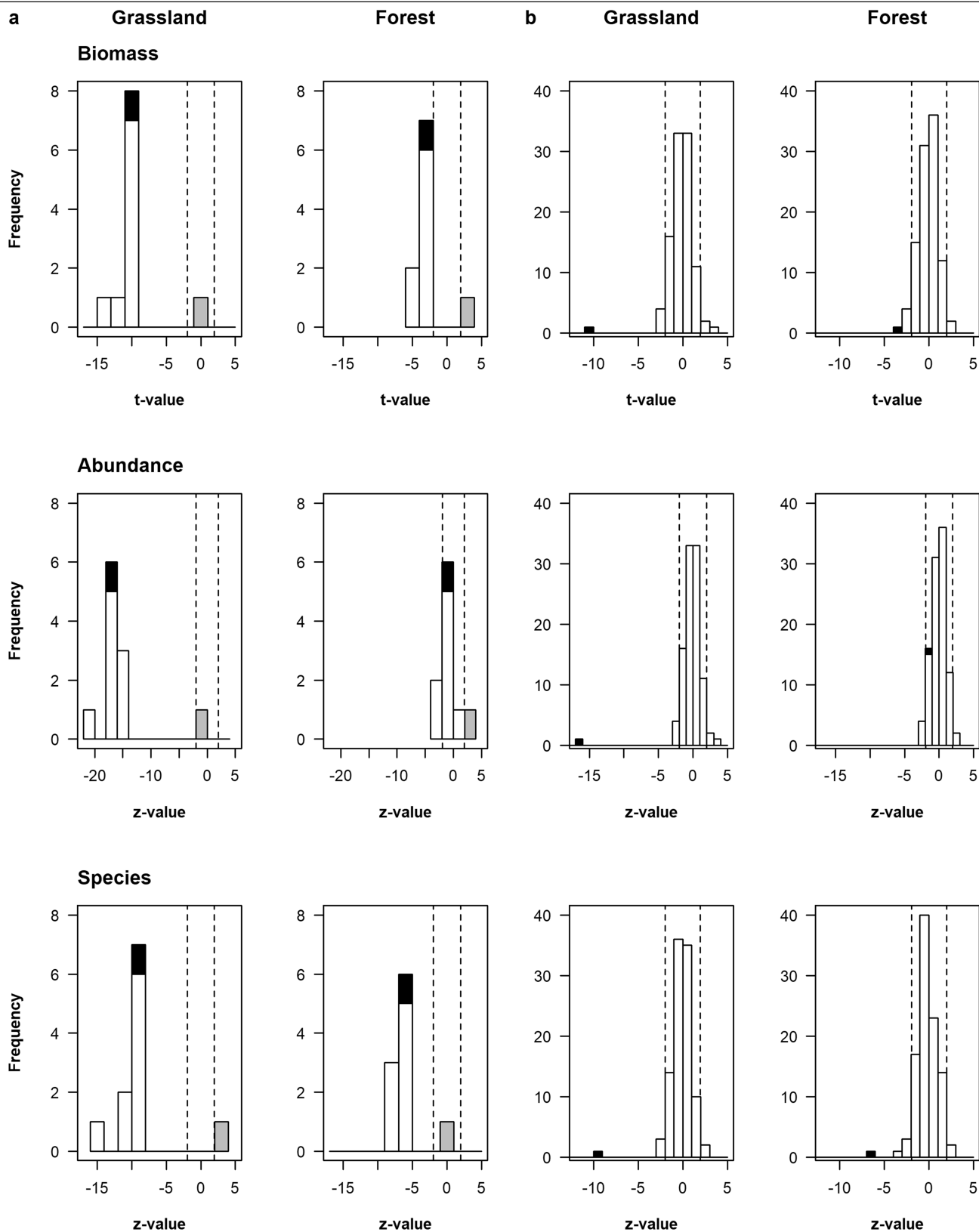
**Peer review information** Nature thanks Simon Leather and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Effects of weather variables on arthropod numbers.** Effects of mean winter temperature (November to February) and precipitation during the growing period (March to October) on biomass, abundance and number of species in arthropod communities in 30 forests (orange) and 150 grasslands (blue) across three regions of Germany. Dots represent raw data aggregated per site and year ( $n = 1,406$  (grassland) or 266 (forest) independent samples). Dotted lines indicate non-significant ( $P \geq 0.05$ ) and solid lines indicate significant effects of weather variables ( $P < 0.05$ ), based on linear mixed models that included year, local and landscape land-use intensity as covariates. Shaded areas represent confidence intervals. The effects of winter temperature and

precipitation differed between forests and grasslands. In grasslands, arthropod numbers increased with increasing winter temperature and with increasing precipitation in the growing period; the effect of precipitation was weaker than the effect of winter temperature, and the effects of both weather variables were weaker than the effect of the year (Supplementary Table 1-1). In forests, arthropod numbers decreased with increasing winter temperature and with increasing precipitation in the growing period; the effects of the two weather variables were similarly strong, but slightly weaker than the effect of the year (Supplementary Table 1-1).



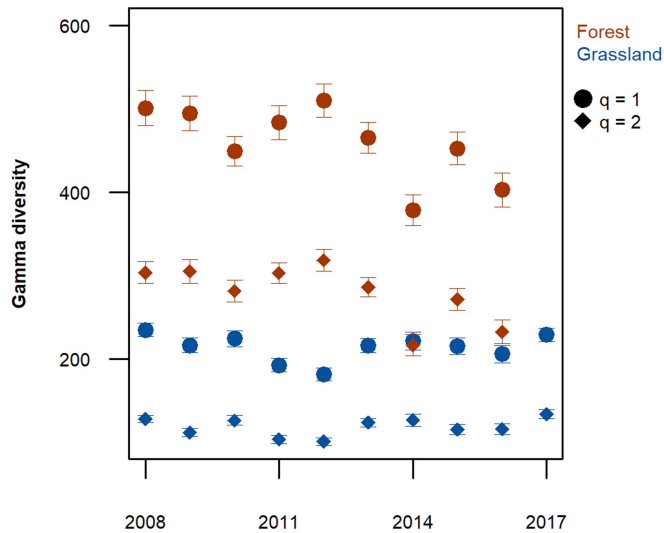
**Extended Data Fig. 2** | See next page for caption.

## Extended Data Fig. 2 | Contribution of individual years to overall trends.

**a.** To assess the contribution of individual years to the overall trend, we repeated the linear mixed models for overall biomass, abundance and number of species, and excluded one year each time. The distribution of  $t$  and  $z$  values for the effect of the year from subset models (white), and from the full models including all years (black), are shown (11 models for grasslands and 10 models for forests). Grey bars denote effect of the year 2008 (the year with the strongest effect on overall trend estimates). **b.** In addition, we tested whether the observed effect of year differed from a random expectation by randomizing the order of years 100× for forests and grasslands before modelling. The distribution of  $t$  and  $z$  values for

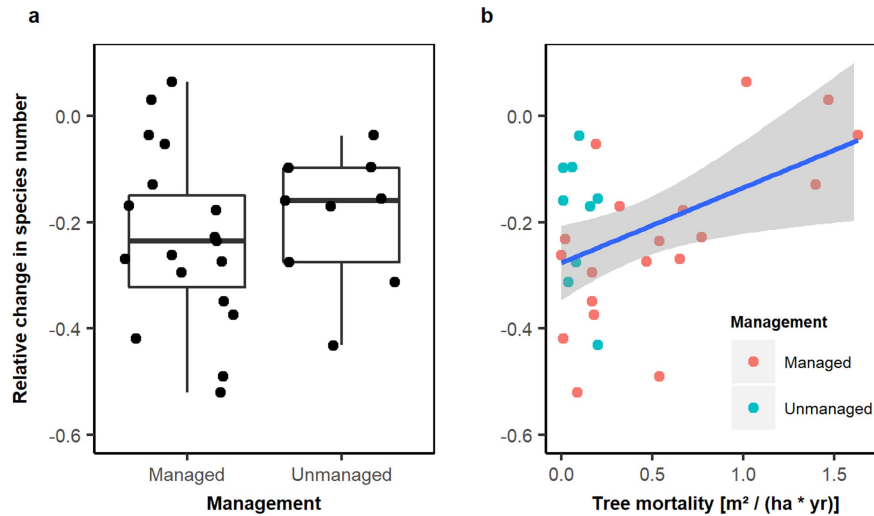
the effect of the year from models with randomly ordered years (white) and models with the years ordered correctly (black) are shown (101 models each for grasslands and forests). Vertical dashed lines indicate levels of significance with  $P < 0.05$ . The results in **a** show that both weaker and stronger temporal trends could be detected when single years were excluded from the analysis, compared to the full model including all years. Results in **b** show that models with the years ordered randomly produced effects of the year that were normally distributed around zero, and only the models with years ordered correctly generated strong temporal trends. For a more detailed discussion, see Supplementary Information section 3.





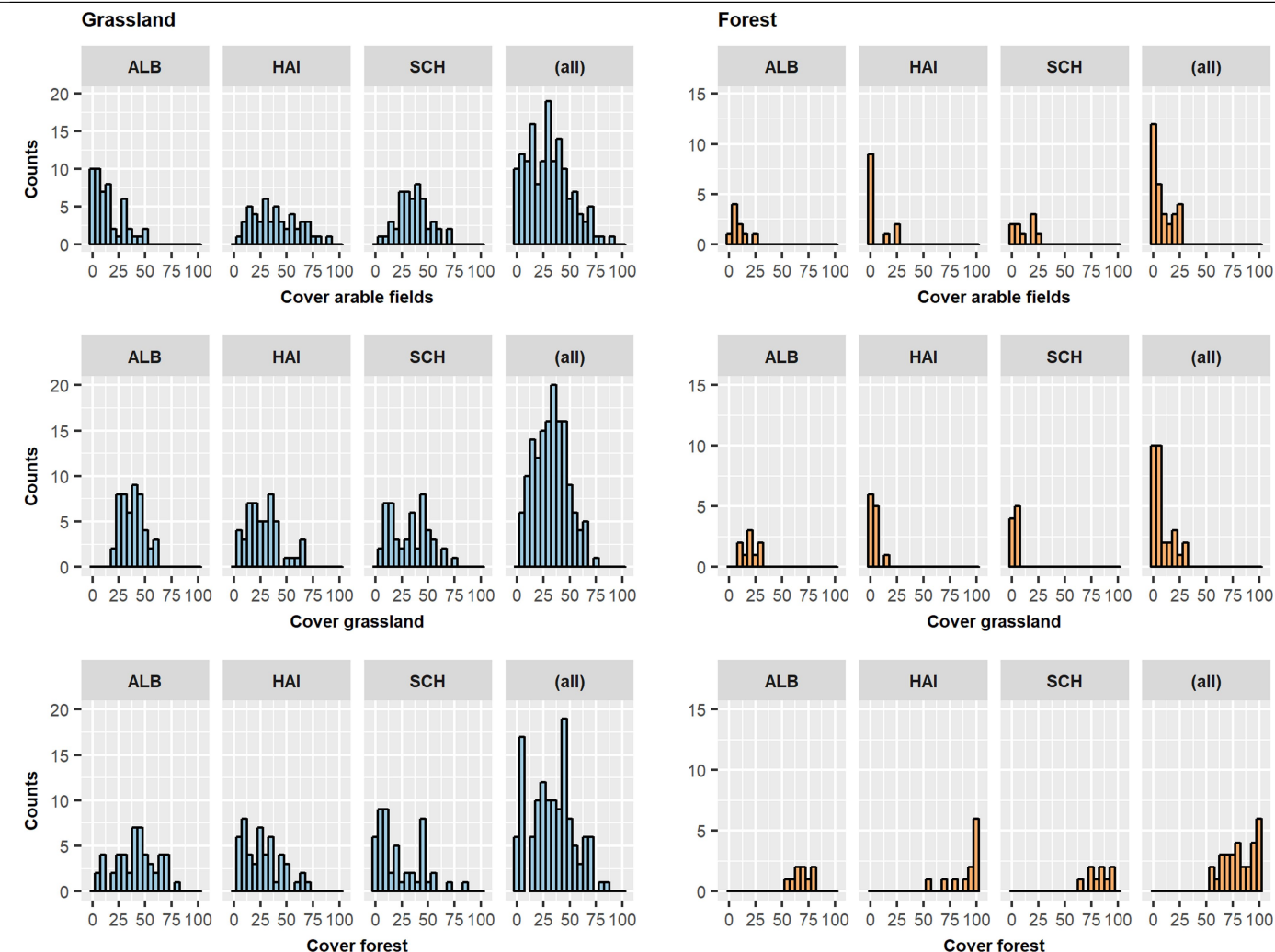
### Extended Data Fig. 3 | Declines in gamma diversity of frequent species.

Estimated gamma diversity (total number of species across all grassland or forest sites) over time. Symbols and error bars shown mean and 95% confidence intervals for gamma diversity, calculated as incidence-based, bias-corrected diversity estimates (Chao's BSS<sup>29</sup>, with 200 bootstrapping runs; Methods) for  $q=1$  and 2 (for  $q=0$ , see Fig. 1). With increasing order  $q$ , the more-frequent species are more strongly weighted ( $q=0$  equals species richness,  $q=1$  equals the exponential of Shannon entropy and  $q=2$  equals the inverse of Simpson diversity; that is, only dominant species affect the diversity measure). This approach accounts for slight differences in site numbers between years caused by limited accessibility or failure of traps. Non-overlapping confidence intervals indicate a significant difference between two sampling years<sup>30</sup>. Figure 1 shows that gamma diversity declines in both forests and grasslands for  $q=0$ . We find that in forests gamma diversity declines when only the more-common species are considered ( $q=1$  and  $q=2$ ), whereas in grasslands there is no overall decline when only the common species are considered. For a more detailed interpretation, see Supplementary Information section 4.



**Extended Data Fig. 4 | Effect of tree mortality on arthropod trends. a,** The relative change in the number of arthropod species between the first two and the final two study years was similar for managed ( $n = 19$ ) and unmanaged ( $n = 9$ ) forest sites ( $z = 0.648$ ,  $P = 0.517$ , derived from a linear mixed model with relative difference in species number as response, harvesting category as fixed and region as random effect). Dots show raw data per site. Boxes represent data within the 25th and 75th percentile, black lines show medians, and whiskers show  $1.5 \times$  the interquartile range. **b,** When considering actual tree mortality between forest inventories in 2009 and 2016, declines in the number of arthropod species were weaker at sites with higher tree mortality ( $z = 2.536$ ,  $P = 0.011$ , derived from a linear mixed model with relative difference in species number as response, tree mortality as fixed and region as random effect). Dots show raw data per site. The blue line visualizes the significant relationship between the change in the number of arthropod species and tree mortality based on the linear mixed model, and the shaded area represents confidence intervals. This suggests that changes in habitat conditions and heterogeneity linked to tree mortality—such

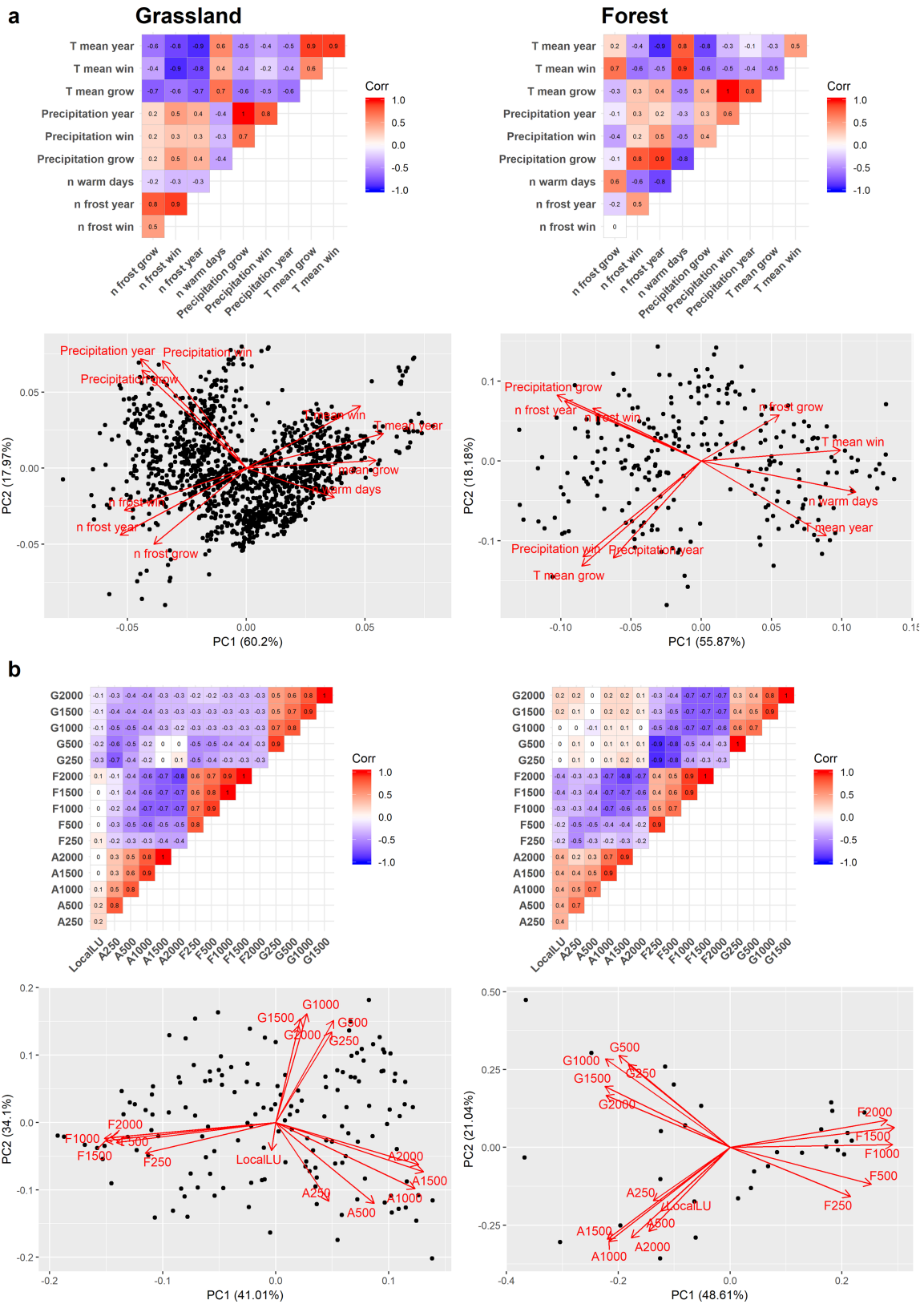
as increasing canopy openness, herb cover or deadwood availability—moderated declines in the number of arthropod species. More research is needed to identify mechanistic relationships. Tree mortality included both natural mortality and timber harvesting. Forest sites had a stand age of, on average, 116 years (minimum of 30 years and maximum of 180 years) and therefore did not include overmature stands. Owing to stand age and because management was abandoned 20 to 70 years before this study started, natural tree mortality was low even in unmanaged stands. We expect increasingly positive effects of natural tree mortality and associated increased structural diversity and heterogeneity<sup>40</sup> on arthropod trends with increasing stand age, but further research is required. In Germany, harvesting is usually conducted as shelterwood cutting. In our sites, the harvested amount over the course of our study reached a maximum of 1% of the standing volume per year. More intense harvesting systems (such as clear cutting), which lead to less heterogeneous habitat conditions, may not have similar moderating effects on arthropod declines.



**Extended Data Fig. 5 | Distribution of landscape-level land-use variables.**

Data distribution of the cover of arable fields, grassland and forest within 1,000 m surrounding each of the 150 grassland and 30 forest sites for each

region, and for all regions in total. ALB, Schwäbische Alb; HAI, Hainich-Dün; SCH, Schorfheide-Chorin.

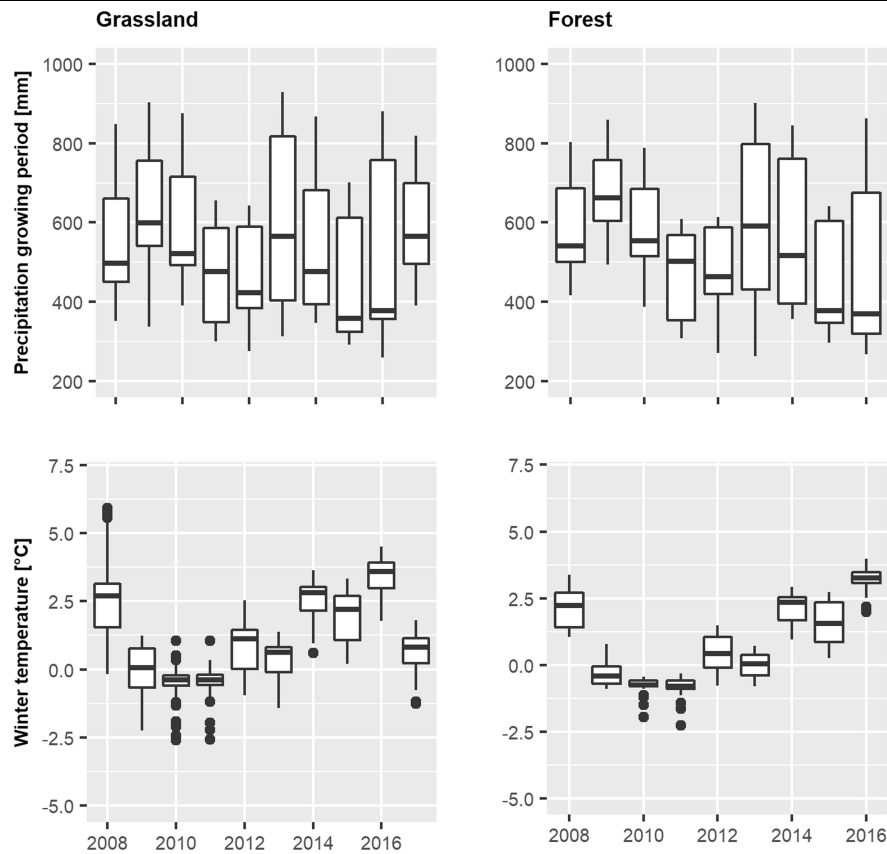


Extended Data Fig. 6 | See next page for caption.



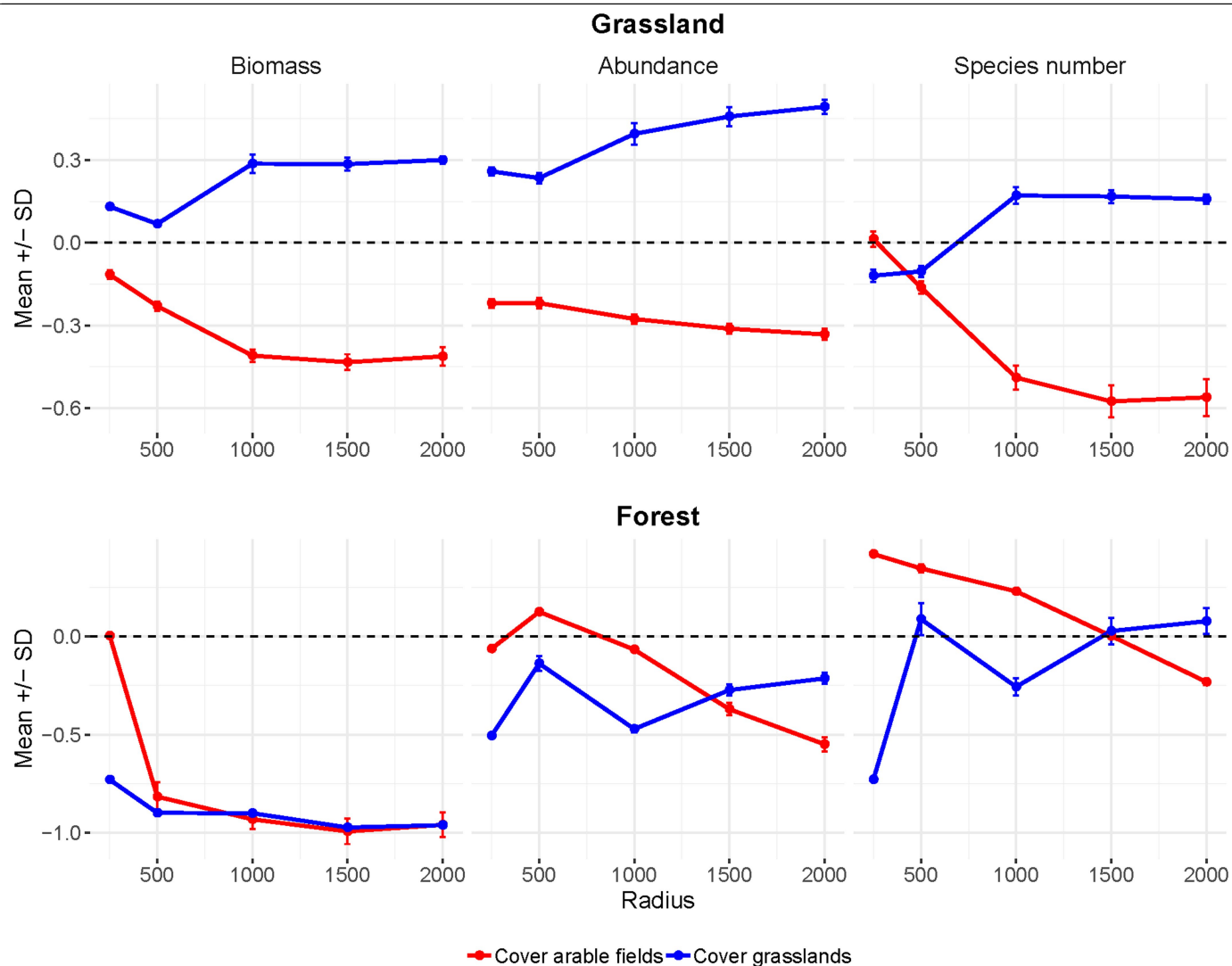
**Extended Data Fig. 6 | Correlations among weather and among land-use variables. a, b.** Coefficients of pairwise correlations and PCAs for weather variables (**a**) and land-use variables (**b**). Temperature-related data are based on observed air temperature by weather stations at each site. Precipitation is derived from gauge-corrected radar observations (RADOLAN, Deutscher Wetterdienst). For each site and year, we calculated mean temperature ( $T_{\text{mean}}$ ), number of frost days (daily minimum temperature  $<0^{\circ}\text{C}$ ;  $n_{\text{frost}}$ ), number of warm days (daily mean temperature  $>20^{\circ}\text{C}$ ;  $n_{\text{warm days}}$ ) and precipitation sum in mm (precipitation) for three different periods: winter (November of the previous year to February; win), growing period (March to October; grow) and

year (November of the previous year to October; year). The number of independent observations for weather variables was  $n = 1,406$  (grasslands) or 266 (forests). Land-use variables include local land-use intensity (local LU) and cover of arable fields (A), grassland (G) and forest (F) at different spatial scales (250, 500, 1,000, 1,500 and 2,000 m). The number of independent observations for land-use variables equalled the number of sites;  $n = 150$  (grasslands) or  $n = 30$  (forests). On the basis of correlations and PCA results, we chose mean winter temperature and precipitation during the growing period, as well as cover of arable fields and cover of grassland, as ecologically meaningful and the least-correlated explanatory variables for modelling arthropod data.



**Extended Data Fig. 7 | Temporal patterns in weather conditions.** Temporal patterns of the sum of precipitation during the growing period (March to October) and mean winter temperature (November of the previous year to February) for 150 grassland and 30 forest sites ( $n=1,406$  (grassland) or 266 (forest) independent observations). Boxes represent data within the 25th and 75th percentile, black lines show medians and whiskers show  $1.5 \times$  the interquartile range. A linear mixed model for each response variable, with year

as a fixed effect and the site nested in the region as a random effect, indicate that winter temperature increased (grassland,  $z=10.90$ ,  $P \leq 0.001$ ; forest,  $z=8.24$ ,  $P \leq 0.001$ ) and precipitation during the growing period decreased during our study period (grassland,  $z=-6.53$ ,  $P \leq 0.001$ ; forest,  $z=-8.44$ ,  $P \leq 0.001$ ). We are currently not able to quantify whether and how much the observed trends in arthropod numbers were affected by changes in climatic conditions (Supplementary Information section 2).



**Extended Data Fig. 8 | Results from multiscale analysis.** Mean and s.d. of Pearson's coefficients of correlation between arthropod numbers (biomass, abundance and number of species) and landscape-level land-use variables (cover of arable fields and cover of grassland) for radii of 250–2,000 m around 150 grassland sites and 30 forest sites. Only data from a random subset of sites with non-overlapping buffers at the 2,000-m scale were used. The randomized subsampling of sites with non-overlapping buffers and the calculation of

correlations was repeated 100 times (median number of sites per subsample was  $n = 18$  (grassland) or 17 (forest)). The 1,000-m scale was used for modelling arthropod numbers for both grassland and forests because (i) the correlation coefficients appeared to plateau at this scale in grasslands, (ii) the range of landscape-level land-use variables at small spatial scales in forests was small and (iii) buffers of neighbouring plots overlapped more extensively at higher spatial scales.

Extended Data Table 1 | Details on arthropod numbers

	Grassland	Forest30	Forest 140	Sum
Abundance				
Identified and unidentified taxa	864,548	80,624	117,731	1,032,279
Unidentified taxa	685,738	NA	NA	685,738
Identified taxa	178,810	80,624	117,731	346,541
Species number	1,309	1,582	1,634	2,675
Biomass[g]	5,637	2,676	4,199	11,642
Weak disperser				
Abundance	29,979	5,744	6,580	
Species number	209	136	148	
Strong disperser				
Abundance	133,710	74,492	110,794	
Species number	946	1,375	1,418	
Carnivores				
Abundance	10,503	9,543	12,611	
Species number	344	531	517	
Omnivores				
Abundance	3,849	21,878	36,900	
Species number	179	354	382	
Herbivores				
Abundance	163,054	20,016	33,252	
Species number	720	295	341	
Myceto-detritivores				
Abundance	1,403	29,158	34,964	
Species number	64	395	394	

Total biomass, number of individual arthropods and number of arthropod species from 150 grassland and 30 or 140 forest sites. Data are available for each year from 2008 to 2017 for all 150 grassland sites, and from 2008 to 2016 for 30 forest sites. In addition, data from 2008, 2011 and 2014 are available for 140 forest sites (including the 30 sites with annual data). Information regarding the abundance of arthropod taxa that were not identified to the species level was collected only in grasslands and not in forests. Classification as a weak or strong disperser was based on morphological and behavioural characteristics (Methods). Owing to missing information, not all species could be assigned to a dispersal or trophic group.



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used to collect data.

Data analysis

Data analyses were conducted using R version 3.5.1 including the packages vegan (function decostand), lme4 (function glmer and lmer), reshape2 (function dcast), BiodiversityR (function rankabundance), stats (prcomp), iNext and ggplot2 for graphics; Climate data was processed using the software TubeDB (<https://environmentalinformatics-marburg.github.io/tubedb>).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data used for analyses are publicly available at the Biodiversity Exploratories Information System (<http://doi.org/10.17616/R32P9Q>) <https://www.bexis.uni-jena.de/PublicData/PublicDataSet.aspx?DatasetId=25786>. Raw data are publicly available from the same repository (IDs: 21969, 22007, 22008, 19686, 20366) or will become publicly available after an embargo period of five years from the end of data assembly to give data owners and collectors time to perform their analysis.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	To disentangle local and landscape-level effects of land use on the temporal trends of arthropod communities in grasslands and forests, we used our arthropod data from the Biodiversity Exploratories research program, including more than 1 million individuals and 2,700 arthropod species. Arthropods were collected annually during the growing period from 2008 to 2017 by standardized sampling at 150 grassland plots and from 2008 to 2016 at 30 forest plots. An additional 118 forest plots were sampled in 2008, 2011 and 2014 to test the overall trend across a larger number of plots. Both grassland and forest plots cover gradients in local land-use intensity. Land-use intensity was quantified in the form of compound indices based on grazing, mowing and fertilization intensity in grasslands, and on recent biomass removal, the proportion of non-native tree species and deadwood origin in forests. To analyze landscape-level effects, we quantified the cover of arable fields, grassland and forest within a 2 km radius around each sampling plot. We modelled temporal trends in biomass, abundance and species number of arthropods and of different dispersal and trophic guilds separately for grasslands and forests, and tested for effects of local and landscape-scale land-use intensity on these trends, accounting for weather conditions and different spatial scales.
Research sample	The sample unit is the arthropod community of a grassland or forest plot within a given year. It is characterized by measures of biomass, abundance and species number. A sample unit is considered to represent arthropod populations at our study sites which measured 50m x 50m in grasslands and 100m x 100 m in forests. Data from different months and traps per plot were pooled per plot and year.
Sampling strategy	Standardized sweep-net sampling along 3 50m transects were conducted in grasslands in June and August. These periods represent the start and end of the peak season for arthropods Central Europe. Sweep-netting is most efficient to sample arthropods in grassland habitats. In forests, we used 2 flight-interception traps per plot which provide a broad range of flying arthropods. Traps were operated during the complete growing season. All samples were sorted to order level in the lab and all groups for which taxonomists were available were identified to species level.  No statistical methods were used to predetermine sample size.
Data collection	Samples were operated in the field and sorted in the lab by trained technicians. Identification was done by expert taxonomists. All people involved are listed in the acknowledgments section.
Timing and spatial scale	Grasslands: annual data collection at all 150 plots in June and August 2008 to 2017 (peak season of arthropods in our study regions) Forests: annual data collection from 2008 to 2016 (April to October) at 30 plots (full growing period represented); in addition, 118 plots were sampled (April to October) in 2008, 2011 and 2014 to test for a larger number of plots whether trends are consistent with results based on annual data from the 30 plots described before
Data exclusions	No data was excluded from the analyses
Reproducibility	Our data were collected as part of a monitoring over several years and which cannot be repeated.
Randomization	Study plots were selected from ~3000 candidate plots. Surveys of initial vegetation and land use were conducted on candidate plots by stratified random sampling to ensure that the selected plots covered the whole range of land-use intensity and to minimize confounding effects of spatial position or soil type.
Blinding	Investigators were not aware of the land-use intensity of the plot where they worked, but they could not otherwise be blinded during data collection and analyses for example with respect to the year a sample came from.
Did the study involve field work?	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No

## Field work, collection and transport

Field conditions	Grasslands: sweep-netting was only conducted when the vegetation was dry and wind speed was low Forests: traps were operated at all weather conditions from April to October
Location	Our data were collected in three German regions: (1) Schwäbische Alb in south-western Germany (420 km <sup>2</sup> , 460–860 m above sea level (a.s.l.)); (2) Hainich-Dün in central Germany (1560 km <sup>2</sup> , 285–550 m a.s.l.); and (3) Schorfheide-Chorin in northeastern Germany (1300 km <sup>2</sup> , 3–140 m a.s.l.).
Access and import/export	Fieldwork permits were issued by the responsible state environmental offices of Baden-Württemberg (Regierungspräsidium

Access and import/export	Tübingen, file number 55-3/8852.15), Thüringen (Thüringer Landesverwaltungsamt, file number 13.4 64233/08-08SDH) and Brandenburg (Landesumweltamt Brandenburg, file number RO7/SOB-0907 ).
Disturbance	Activity of investigators was spatially limited to three 50m transects for sweep-netting and short paths to access the flight-interception traps

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	No laboratory animals were involved in the study.
Wild animals	Arthropods of about 2700 different species were collected in the field and killed on-site using CuSO <sub>4</sub> -solution for flight-interceptions traps and ethanol for sweep-netting. Identification of arthropods requires killing and transport to the lab where microscopes can be used.
Field-collected samples	Samples were stored in 93% ethanol at 7°C except for short time periods during transport, sorting and identification.
Ethics oversight	It could not be ruled out that threatened or protected arthropod species would be collected and killed. Thus, permission was required from the authorities which was granted for scientific reasons. These permits were issued by the responsible state environmental offices of Baden-Württemberg (Regierungspräsidium Tübingen, file number 55-3/8852.15), Thüringen (Thüringer Landesverwaltungsamt, file number 13.4 64233/08-08SDH) and Brandenburg (Landesumweltamt Brandenburg, file number RO7/SOB-0907 ).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Evolution of the new head by gradual acquisition of neural crest regulatory circuits

<https://doi.org/10.1038/s41586-019-1691-4>

Received: 22 October 2018

Accepted: 26 September 2019

Published online: 23 October 2019

Megan L. Martik<sup>1</sup>, Shashank Gandhi<sup>1</sup>, Benjamin R. Uy<sup>1</sup>, J. Andrew Gillis<sup>2,3</sup>, Stephen A. Green<sup>1</sup>, Marcos Simoes-Costa<sup>4</sup> & Marianne E. Bronner<sup>1\*</sup>

The neural crest, an embryonic stem-cell population, is a vertebrate innovation that has been proposed to be a key component of the ‘new head’, which imbued vertebrates with predatory behaviour<sup>1,2</sup>. Here, to investigate how the evolution of neural crest cells affected the vertebrate body plan, we examined the molecular circuits that control neural crest development along the anteroposterior axis of a jawless vertebrate, the sea lamprey. Gene expression analysis showed that the cranial subpopulation of the neural crest of the lamprey lacks most components of a transcriptional circuit that is specific to the cranial neural crest in amniotes and confers the ability to form craniofacial cartilage onto non-cranial neural crest subpopulations<sup>3</sup>. Consistent with this, hierarchical clustering analysis revealed that the transcriptional profile of the lamprey cranial neural crest is more similar to the trunk neural crest of amniotes. Notably, analysis of the cranial neural crest in little skate and zebrafish embryos demonstrated that the transcriptional circuit that is specific to the cranial neural crest emerged via the gradual addition of network components to the neural crest of gnathostomes, which subsequently became restricted to the cephalic region. Our results indicate that the ancestral neural crest at the base of the vertebrate lineage possessed a trunk-like identity. We propose that the emergence of the cranial neural crest, by progressive assembly of an axial-specific regulatory circuit, allowed the elaboration of the new head during vertebrate evolution.

Gans and Northcutt’s ‘new head’ hypothesis proposed that the emergence of the vertebrate lineage was accompanied by the advent of the neural crest (NC), a stem-cell population that arises within the forming central nervous system (CNS) in all vertebrates<sup>1,2</sup>. These cells subsequently leave the CNS, migrate to diverse locations and differentiate into many tissues, including peripheral ganglia and the craniofacial skeleton<sup>4,5</sup>. As vertebrates evolved, NC cells contributed to morphological novelties such as jaws, which enabled the expansion of vertebrates.

It has been proposed that a pan-vertebrate NC gene regulatory network (GRN), which invokes sequential deployment of signalling and transcriptional events, underlies the formation of this cell type. The core of the NC GRN is largely conserved across vertebrates, including the sea lamprey *Petromyzon marinus*, a jawless (cyclostome) vertebrate, and has been primarily studied at cranial levels. However, key transcription factors such as *Ets1* and *Twist* are deployed later in the lamprey GRN than in amniotes<sup>6,7</sup>, suggesting that there are regulatory differences between cyclostomes and gnathostomes. Furthermore, some NC derivatives are unique to gnathostomes, including jaws at cranial levels, a vagal-derived enteric nervous system, and sympathetic ganglia at trunk levels<sup>8,9</sup>. This raises the possibility that network differences in axial regionalization

of the neural crest may have contributed to the presence of these gnathostome cell types.

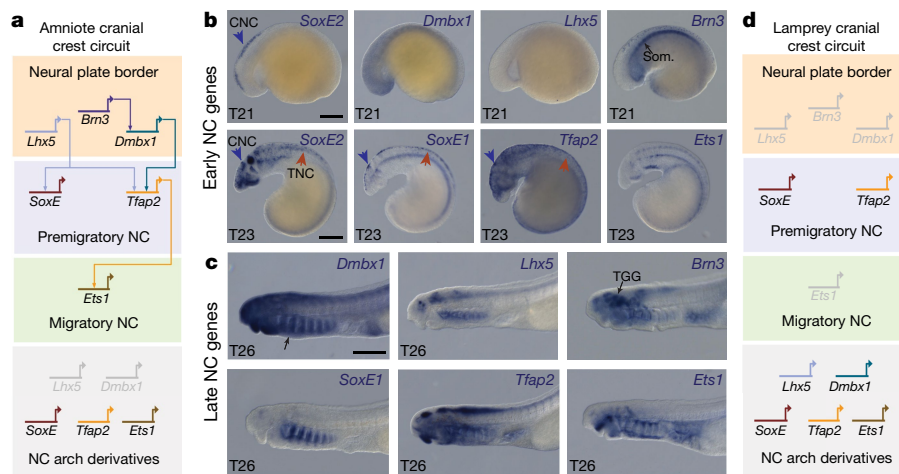
In jawed vertebrates, the NC is subdivided along the body axis into cranial, vagal, and trunk subpopulations<sup>5</sup>. By contrast, lampreys lack an intermediate vagal population, suggesting that there are two major subpopulations: cranial and trunk<sup>8</sup>. It is unclear how axial identity in the lamprey is controlled at a molecular level. Avian embryos possess a ‘cranial crest-specific’ NC GRN subcircuit that can drive the differentiation of trunk NC into ectomesenchymal derivatives<sup>3</sup>. In this subcircuit, the transcription factors *Brn3c*, *Lhx5*, and *Dmbx1* are expressed at the neural plate border and, in turn, activate the expression of *Sox8* in the premigratory cranial NC and *Ets1* in the migratory cranial NC (Fig. 1a). By contrast, the pan-NC genes *Tfap2b* and *Sox10* are expressed all along the body axis<sup>10</sup>.

## Lamprey lacks full cranial NC subcircuit

We investigated whether this cranial subcircuit is a general feature of vertebrates by examining whether lampreys possess a homologous spatiotemporal regulatory state. Taking a candidate approach, we analysed the expression of cranial crest subcircuit orthologues in lamprey

<sup>1</sup>Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. <sup>2</sup>Department of Zoology, University of Cambridge, Cambridge, UK. <sup>3</sup>Whitman Center, Marine Biological Laboratory, Woods Hole, MA, USA. <sup>4</sup>Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. \*e-mail: mbronner@caltech.edu





**Fig. 1 | Lamprey cranial NC lacks most components of a chick ‘cranial crest circuit’.** **a**, Biotapestry model of cranial NC-specific gene regulatory circuit that drives skeletal differentiation in amniotes. **b**, Expression of lamprey orthologues of amniote cranial NC-specific genes at T21 and T23. Blue arrows, expression in the cranial neural crest (CNC); red arrows, expression in the trunk neural crest (TNC). **c**, Late expression of cranial NC-specific orthologues in pharyngeal arch NC derivatives (black arrow). **d**, Biotapestry model of the lamprey circuit with the addition of late module expression of markers in pharyngeal arch NC derivatives. TGG, trigeminal ganglia. Scale bars, 250  $\mu$ m. Reproducible on  $n \geq 5$  embryos per time point for  $n \geq 10$  experiments.

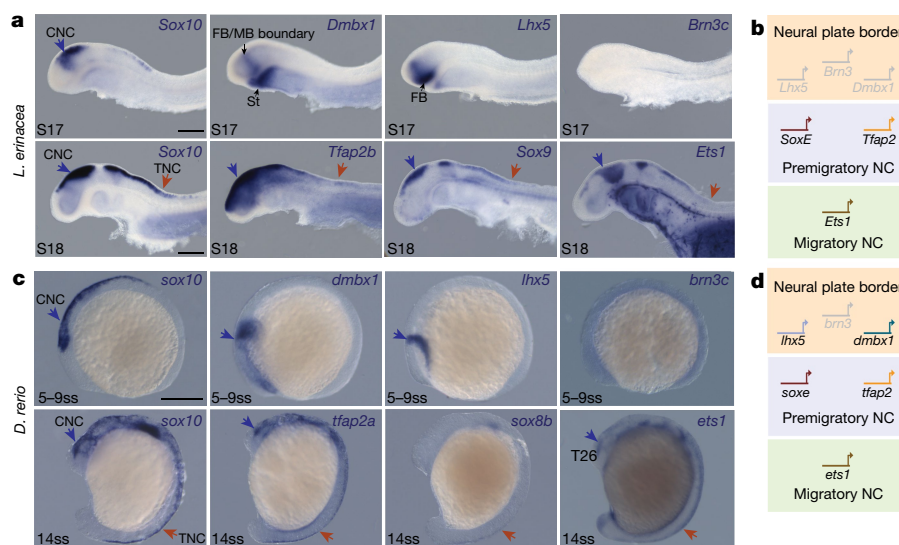
embryos at different developmental stages. In contrast to amniotes, in lampreys the premigratory and migratory NC did not appear to express *Bmx3*, *Lhx5*, *Dmbx1*, or *Ets1* (Fig. 1b). The lack of most cranial-specific regulatory factors suggests a high degree of divergence between the early regulatory states of the lamprey and amniote NC. By contrast, *SoxE1* and *Tfap2* were robustly expressed in premigratory and migratory NC along the entire lamprey body axis (Fig. 1b, Extended Data Fig. 1a–d). No *SoxE* family member was restricted in expression to the cranial NC, as *SoxE8* is in amniotes. Of note, lamprey *SoxE* transcription factors are homologous to gnathostome *SoxE8*, *SoxE9* and *SoxE10*, and *SoxE* paralogue usage varies across gnathostomes<sup>11–13</sup>. Consistent with the lack of restricted ‘cranial-specific’ expression, ectomesenchymal derivatives have been found at trunk levels in the lamprey dorsal fin<sup>9</sup>.

How then did this regulatory subcircuit evolve? Notably, genes from the amniote cranial crest subcircuit are present in the lamprey genome

but are expressed later in pharyngeal arches populated by NC cells (Fig. 1c, d, Extended Data Figs. 1e–l, 2). An intriguing possibility is that these genes were expressed only in late NC derivatives of early vertebrates and were gradually co-opted to earlier developmental stages in gnathostomes. According to this scenario, genes involved in NC differentiation in early vertebrates were co-opted to the specification program of gnathostomes at all axial levels. With subsequent regulatory modifications, they became cranially restricted, possibly endowing the cranial NC with novel morphogenetic features while the trunk NC lost the ability to make cranial-like derivatives<sup>14,15</sup>.

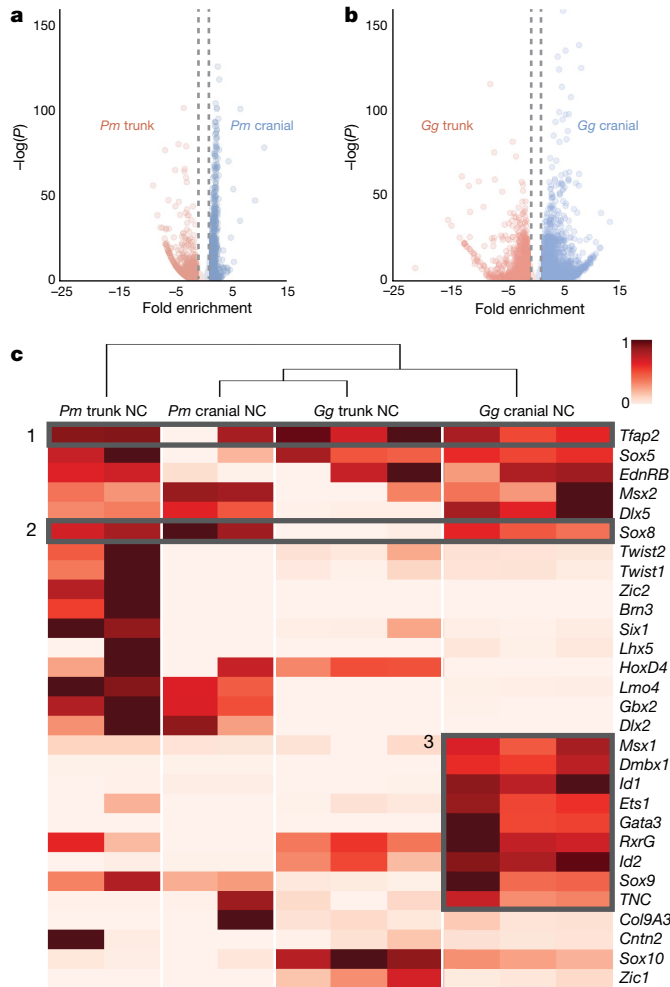
### Skates add *Ets1* to cranial NC subcircuit

To explore the transcriptional landscape in a basal jawed vertebrate, we examined candidate elements of the cranial NC subcircuit in the little



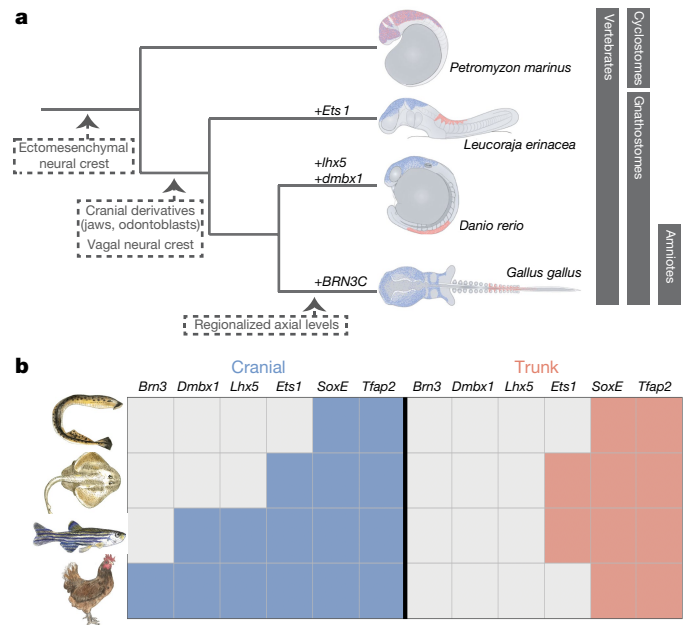
**Fig. 2 | Nodes of an early cranial NC-specific circuit were acquired in the cranial NC progressively throughout gnathostome evolution.** **a**, Expression of cranial NC-specific orthologues in the little skate at stages (S) 17 and 18. Blue arrow, expression of orthologues in the cranial NC; red arrows, expression of orthologues in the trunk NC. **b**, Biotapestry model of the skate circuit with the addition of a novel node, *Ets1*. **c**, Expression of cranial NC-specific orthologues in the zebrafish at 5–9 somite stage (ss) and 14ss. **d**, Biotapestry model of the zebrafish circuit with the addition of novel early nodes, *lhx5* and *dmbx1*. FB/MB, forebrain/midbrain; St, stomodeum. Scale bars, 250  $\mu$ m. For skates, in situ hybridizations were reproducible on  $n \geq 2$  embryos for  $n \geq 2$  experiments. For zebrafish, in situ hybridizations were reproducible on  $n \geq 5$  embryos per time point for  $n \geq 10$  experiments.

the zebrafish at 5–9 somite stage (ss) and 14ss. **d**, Biotapestry model of the zebrafish circuit with the addition of novel early nodes, *lhx5* and *dmbx1*. FB/MB, forebrain/midbrain; St, stomodeum. Scale bars, 250  $\mu$ m. For skates, in situ hybridizations were reproducible on  $n \geq 2$  embryos for  $n \geq 2$  experiments. For zebrafish, in situ hybridizations were reproducible on  $n \geq 5$  embryos per time point for  $n \geq 10$  experiments.



**Fig. 3 | Tissue-specific RNA sequencing comparisons between lamprey and chicken reveal that the ancestral NC had a more trunk-like identity.** **a**, Volcano plot showing lamprey differential enrichments of cranial (blue) and trunk (red) genes by population RNA sequencing (RNA-seq; 100 embryos were dissected for each of  $n = 2$  biological replicates; adjusted  $P < 0.05$ ). **b**, Volcano plot showing enrichment of genes in the cranial (blue) versus the trunk (red) NC in chicken ( $\geq 15$  heads and 5 trunks were dissected and prepared for FACS for each of  $n = 3$  biological replicates; adjusted  $P < 0.05$ ). **c**, Hierarchical clustering analysis of all RNA-seq libraries focused on the NC GRN reveals similarities and differences between axial levels among species.

skate *Leucoraja erinacea*, a member of a chondrichthyan gnathostome outgroup to the bony fishes. *Sox9* and *Sox10* of the SoxE genes, as well as *Tfap2b* and *Ets1*, were expressed at all axial levels and not restricted to the cranial crest (Fig. 2a, Extended Data Fig. 3). Because *Ets1* appears in the little skate migratory NC, as in other gnathostomes, we conclude that this early node was a novelty acquired by the NC GRN before the divergence of cartilaginous and bony fishes (Fig. 2a, b, Extended Data Fig. 3). Later, after NC cells had migrated to and populated the pharyngeal arches, *SoxE*, *Tfap2b*, and *Ets1* were expressed within the arches (Extended Data Fig. 4). Trunk NC cells in the little skate give rise to ectomesenchymal dermal denticles ('cranial-like' derivatives), consistent with our observation that cranial subcircuit genes in the little skate are not restricted to the head but can drive the differentiation of skeletogenic derivatives in the trunk<sup>16</sup>. In the fossil record, many stem-gnathostomes possessed extensive dermal armour, which has been retained, albeit with the dental component reduced and modified, within the gnathostome crown group (for example, the dermal denticles of chondrichthyans and the dentinous scales of *Polypterus* and coelacanth). Thus, dental tissues in the postcranial dermal skeleton appear to be ancestral for gnathostomes.



**Fig. 4 | Model of the evolution of NC axial levels during vertebrate evolution.** **a**, Our data suggest that the ancestral NC was a more uniform population of cells along the body axis that underwent gradual regulatory modifications during gnathostome evolution. **b**, Progressive restriction of the 'cranial circuit genes' to the cranial axial level led to axial specialization of the NC regulatory program.

### Further elaboration of the NC subcircuit

Notably, in the teleost *Danio rerio* (zebrafish), *lhx5* and *dmbx1* are expressed in the early cranial circuit but not in later pharyngeal arch derivatives (Fig. 2c, d, Extended Data Figs. 5, 6). In addition, *sox8b*, *sox10*, *tfap2a*, and *ets1*, but not *brn3c* (also known as *pou4f3*) are expressed in premigratory and migratory crest at all axial levels (Fig. 2c, Extended Data Fig. 5). Rather than being restricted to the cranial NC, many of these factors are also expressed in the zebrafish trunk, raising the possibility that the resolution of axial level potential may have arisen within sarcopterygians. Furthermore, in situ analysis of pharyngeal arch derivatives in both little skate and zebrafish lend support to temporal shifts of cranial specific regulatory nodes from later NC derivatives to an early specification program (Extended Data Figs. 4, 6). With loss of nodes from late derivatives and addition to an earlier program, this suggests that regulatory modifications arose gradually throughout gnathostome evolution (Extended Data Fig. 1m).

### Lamprey cranial NC is more trunk-like

Our candidate gene approach suggests that extensive changes occurred in the NC regulatory state between jawless and jawed vertebrates. To investigate further, we conducted a comparative transcriptome analysis of cranial and trunk NC subpopulations in lamprey and chicken (Fig. 3). We obtained premigratory lamprey NC by micro-dissecting segments of cranial and trunk dorsal neural tubes at Tahara stages T21 and T23.5. For chick, we isolated premigratory NC populations using enhancers that drive eGFP expression in cranial or trunk neural crest populations for fluorescence-activated cell sorting (FACS) at Hamburger–Hamilton stages HH9+ and HH18, respectively. After cDNA library preparation and sequencing, differential expression analysis showed that far fewer genes were significantly enriched in lamprey cranial NC versus trunk NC, compared with chick cranial NC versus trunk NC (1,233 genes in lamprey compared with 2,794 in chicken; Fig. 3a, b).

To better understand how each library correlated to the others in an unbiased fashion, we mapped each to a common reference

transcriptome, created by aligning proteomes using BLAT<sup>17</sup> and compiling matching sequences as a consensus alignment between species for Bowtie mapping (Supplementary Table 1). We next performed hierarchical clustering analysis of all known NC GRN genes (Fig. 3c). Consistent with our previous *in situ* hybridization analysis (Fig. 1b), *Tfap2* (known as *TFAP2A* in chicken) was enriched at all axial levels in both chicken and lamprey, *SOX8* was enriched in chicken cranial NC but in both cranial and trunk lamprey populations, and *DMBX1* and *ETS1* were enriched in chick but not lamprey cranial datasets (Fig. 3c).

Notably, we found that lamprey cranial NC populations correlated more closely to chicken trunk than lamprey trunk libraries, suggesting that basal NC was ‘trunk-like’ in its regulatory program (Fig. 3c). These results suggest that cyclostomes possess a simpler and more trunk-like cranial crest GRN, with potentially important implications for the evolution of NC subpopulations (Fig. 4a). Accordingly, we speculate that the ancestral neural crest may have been relatively homogeneous and trunk-like. Throughout evolution of the vertebrate lineage, we propose that key transcription factors were progressively co-opted into an early, cranially restricted circuit, whereas some features, such as skeletogenic potential, were lost from the trunk.

## Conclusions

The differences in expression of axial level-specific genes contrast with the deep conservation of the pan-NC program<sup>6,18</sup>. Transcription factors such as SoxE genes, *Tfap2*, and *Id* may be the rudiments of a larger, more complex cranial crest GRN that was expanded during early vertebrate evolution with the incorporation of genes such as *Dmbx1*, *Brn3c*, *Ets1*, and *Lhx5*. Consistent with these findings, the basal chordate amphioxus lacks expression at the neural plate border of genes such as *Dmbx*, *Brn3*, and *Ets*, as well as core NC genes such as *SoxE*, *FoxD*, *Tfap2*, and *Id*, although these genes are expressed in other tissues<sup>19–22</sup>. Our observations also show that some of these ‘novel’ genes are expressed at later stages of NC formation in lamprey, consistent with the possibility that elaboration of the GRN might have involved the co-option of parts of differentiation programs to earlier portions of the network, perhaps by acquisition of new regulatory elements responsible for their heterochronic shift<sup>23</sup>. Thus, the pan-NC program was probably the ancestral molecular recipe to make the NC, with the subsequent elaboration of axial level-specific regulatory programs conferring important differences in developmental potential along the body axis. Given that many key NC derivatives are gnathostome innovations, we hypothesize these derivatives may have been gained as the result of gene regulatory differences associated with axial level-specific regulatory programs.

Together, our results suggest the following scenario to explain the evolution of NC subpopulations (Fig. 4). We suggest that the NC of early vertebrates was uniform and similar to amniote trunk populations, and that the division of the NC into cranial and trunk subpopulations occurred early in vertebrate evolution (Fig. 4a). Consistent with evolutionary expansion of NC cells in the vertebrate lineage, our molecular analysis of the cranial NC reveals unexpected differences between lampreys and their gnathostome counterparts (Fig. 4b). Given that the Hox code was already linked to segmentation of the CNS in basal vertebrates, posteriorizing influences of Hox genes and other factors may be sufficient to account for the subtle transcriptional differences observed between these two populations<sup>24,25</sup>. We cannot rule out the possibility that cyclostomes lost NC subpopulations during the course of evolution. However, the relative scarcity of cranial NC-specific subcircuit factors in the lamprey cranial crest might suggest that the gnathostome cranial NC GRN has undergone extensive elaboration from a regulatory standpoint. Thus, we propose that regionalization of the NC, with both emergence of new subpopulations and expansion of the cranial crest GRN, was a key element for driving the evolution and expansion of gnathostomes during vertebrate evolution.

What does this mean for the new head hypothesis? We posit that the NC component of the new head, rather than arising *in toto* at the base of the vertebrate lineage, underwent continued regulatory modifications, evolving gradually during the course of vertebrate evolution. Our data suggest that early vertebrates possessed a relatively simple NC that initially arose as a fairly uniform population along the body axis and lacked region-restricted regulatory programming. During gnathostome evolution, the cranial NC appears to have gained regulatory complexity that modulated its differentiation capacity, gaining some individual cell fates while restricting others. We propose that co-option of distinct genes into a cranial NC-specific module enabled this progressive specialization of NC regulatory programs, leading to the specific axial subpopulations and morphological novelties of the gnathostome body plan.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1691-4>.

- Gans, C. & Northcutt, R. G. Neural crest and the origin of vertebrates: a new head. *Science* **220**, 268–273 (1983).
- Glenn Northcutt, R. The new head hypothesis revisited. *J. Exp. Zool. B Mol. Dev. Evol.* **304**, 274–297 (2005).
- Simoes-Costa, M. & Bronner, M. E. Reprogramming of avian neural crest axial identity and cell fate. *Science* **352**, 1570–1573 (2016).
- Le Douarin, N. M. & Smith, J. Development of the peripheral nervous system from the neural crest. *Annu. Rev. Cell Biol.* **4**, 375–404 (1988).
- Le Douarin, N. *The Neural Crest* (Cambridge Univ. Press, 1982).
- Sauka-Spengler, T., Meulemans, D., Jones, M. & Bronner-Fraser, M. Ancient evolutionary origin of the neural crest gene regulatory network. *Dev. Cell* **13**, 405–420 (2007).
- Nikitina, N., Sauka-Spengler, T. & Bronner-Fraser, M. Dissecting early regulatory relationships in the lamprey neural crest gene network. *Proc. Natl Acad. Sci. USA* **105**, 20083–20088 (2008).
- Green, S. A., Uy, B. R. & Bronner, M. E. Ancient evolutionary origin of vertebrate enteric neurons from trunk-derived neural crest. *Nature* **544**, 88–91 (2017).
- Häming, D. et al. Expression of sympathetic nervous system genes in Lamprey suggests their recruitment for specification of a new vertebrate feature. *PLoS One* **6**, e26543 (2011).
- Betancur, P., Bronner-Fraser, M. & Sauka-Spengler, T. Genomic code for *Sox10* activation reveals a key regulatory enhancer for cranial neural crest. *Proc. Natl Acad. Sci. USA* **107**, 3570–3575 (2010).
- Haldin, C. E. & LaBonne, C. SoxE factors as multifunctional neural crest regulatory factors. *Int. J. Biochem. Cell Biol.* **42**, 441–444 (2010).
- McCauley, D. W. & Bronner-Fraser, M. Importance of SoxE in neural crest development and the evolution of the pharynx. *Nature* **441**, 750–752 (2006).
- Lee, E. M. et al. Functional constraints on SoxE proteins in neural crest development: the importance of differential expression for evolution of protein activity. *Dev. Biol.* **418**, 166–178 (2016).
- Green, S. A., Simoes-Costa, M. & Bronner, M. E. Evolution of vertebrates as viewed from the crest. *Nature* **520**, 474–482 (2015).
- Martik, M. L. & Bronner, M. E. Regulatory logic underlying diversification of the neural crest. *Trends Genet.* **33**, 715–727 (2017).
- Gillis, J. A., Alsema, E. C. & Criswell, K. E. Trunk neural crest origin of dermal denticles in a cartilaginous fish. *Proc. Natl Acad. Sci. USA* **114**, 13200–13205 (2017).
- Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
- Sauka-Spengler, T. & Bronner-Fraser, M. Evolution of the neural crest viewed from a gene regulatory perspective. *Genesis* **46**, 673–682 (2008).
- Wada, H., Kobayashi, M. & Zhang, S. *Ets* identified as a *trans*-regulatory factor of amphioxus *Hox2* by transgenic analysis using ascidian embryos. *Dev. Biol.* **285**, 524–532 (2005).
- Meulemans, D. & Bronner-Fraser, M. Amphioxus and lamprey AP-2 genes: implications for neural crest evolution and migration patterns. *Development* **129**, 4953–4962 (2002).
- Takahashi, T. & Holland, P. W. H. Amphioxus and ascidian *Dmbx* homeobox genes give clues to the vertebrate origins of midbrain development. *Development* **131**, 3285–3294 (2004).
- Yu, J.-K., Meulemans, D., McKeown, S. J. & Bronner-Fraser, M. Insights from the amphioxus genome on the origin of vertebrate neural crest. *Genome Res.* **18**, 1127–1132 (2008).
- Davidson, E. H. & Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–800 (2006).
- Parker, H. J., Bronner, M. E. & Krumlauf, R. A Hox regulatory network of hindbrain segmentation is conserved to the base of vertebrates. *Nature* **514**, 490–493 (2014).
- Parker, H. J., Bronner, M. E. & Krumlauf, R. The vertebrate Hox gene regulatory network for hindbrain segmentation: evolution and diversification. *BioEssays* **38**, 526–538 (2016).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### Animal husbandry and embryo collection

Adult sea lamprey were obtained from the US Fish and Wildlife Service and Department of the Interior. Embryos were cultured according to previously published protocols and staged according to Tahara staging methods<sup>12,26</sup>. All lamprey embryology work was completed in compliance with California Institute of Technology Institutional Animal Care and Use Committee (IACUC) protocol 1436. Skate eggs were obtained by the Marine Biological Laboratory (MBL) Marine Resource Center, and embryos were cultured as previously described<sup>16</sup>. All skate embryology work was compliant with animal protocols approved by the IACUC at the MBL. Adult zebrafish were maintained in the Beckman Institute Fish Facility at Caltech, and all animal and embryo work was compliant under approved IACUC protocol 1346. Fertilized chicken eggs were obtained from a local farm in Sylmar, CA. No statistical methods were used to predetermine sample size for analyses. For in situ hybridization, embryos were pooled from different breeding pairs (fish), brooding stocks (skates), or embryo batches (lamprey) to ensure replication of results in multiple fixed collections.

### Cloning of lamprey, skate, and zebrafish orthologues

RNA was extracted from desired embryo stages using the RNAqueous-Micro Kit (Thermo Fisher Scientific), and cDNA was synthesized using a SuperScript III Reverse Transcriptase Kit (Invitrogen). The following gene-specific primers were used to amplify probe template sequences (accession numbers in parentheses): *PmTfap2* (MN410935): F: 5'-GCATCGCGACAGTTGTTTGCTG-3'; R: 5'-GATGCTGTGGTGCCTAATCC-3'; *PmSox2* (MN410934): F: 5'-CGAGTCACTGGATCTGCTGC-3'; R: 5'-CCGTCCAGCACTTGACTCACG-3'; *PmSox1* (MN410933): F: 5'-CGGGCTGAGTCACTTCTCGCATCG-3'; R: 5'-CTCTCGTCTGCTGCGAAGC-3'; *PmEts1a* (SIMRbase: PMZ-0040201): F: 5'-GGACCTTCAAGGAGTACATGAGC-3'; R: 5'-GAGAGCGGTACTCGTGGAAAGT-3'; *PmDmbx1* (Ensembl: ENSPMAG00000008114): F: 5'-GCGCATGAATACCGCCGTCG-3'; R: 5'-TTGCTTTGATGCTGTACAAAG-3'; *PmLhx5* (MN410936): F: 5'-CGTGGCTTCGTGACCCCATC-3'; R: 5'-GAGGCCAGGTAGTCTCTCTTG-3'; *PmBrn3* (SIMRbase: PMZ-0005302): F: 5'-CGAGTCTCTTAACGCGTTAGCTC-3'; R: 5'-GCTCTGGTGGGAGACAATATCCACG-3'; *LeTfap2b* (MN410937): F: 5'-TCCCACTTCCACAGAAGAAT-3'; R: 5'-TCCTGTCTCCAGTTTGGTG-3'; *LeSox10* (MN410938): F: 5'-ACCCCGTTCTGTGTGTCT-3'; R: 5'-GGCAGGTACTGGTCAACTC-3'; *LeSox9* (MN410939): F: 5'-CCCAGCCACTACAATGAGCAG-3'; R: 5'-CCGTACGGCATCAGCAAATG-3'; *LeSox8* (MN410944): F: 5'-CAACTCCGCCCACTACTCC-3'; R: 5'-TGGCGTAGTCAAGGTTGTGTAG-3'; *LeEts1* (MN410940): F: 5'-TTCAGCTGAGAAGCTGGAC-3'; R: 5'-GCAAGACTGTCCGTACAGGAG-3'; *LeDmbx1* (MN410941): F: 5'-CAATCAACACGACAGGACA-3'; R: 5'-GTAAGCTGTCAAGCCCCAGA-3'; *LeLhx5* (MN410942): F: 5'-TCATCGACGAAAACAAATTTGTGTG-3'; R: 5'-TGAATAACCCGCATGTTGAGGC-3'; *LeBrn3c* (MN410943): F: 5'-CTTCAAGCCGACATCACCTAC-3'; R: 5'-TAGATCCCTGCTTGTTCCTGC-3'; *Drtfap2a* (NM\_176859): F: 5'-GTCACGGCATTGATACTGGACTC-3'; R: 5'-TCATTGGCACACTGCTTTACTGAT-3'; *Drsox10* (NM\_131875): F: 5'-GTGAAACACACTTCCCTGGGGATAC-3'; R: 5'-GTGGAGACATGTGTGTATGGCGTC-3'; *Drsox8b* (NM\_001025465): F: 5'-ATGACGAGGAGCGGAAAGTG-3'; R: 5'-GGGTCTGGACAGTGGTGAGAC-3'; *DreIsl* (NM\_001017558): F: 5'-CAGACACGGATCTTGTGAGGGA-3'; R: 5'-CAGTCCAGCTGATGAAGGACTGG-3'; *Drdmbx1a* (NM\_152977): F: 5'-CGTGCCAGTCTACTATCAGTCTC-3'; R: 5'-CTGCTGTGTAGTGCATGCAACC-3'; *Drlhx5* (NM\_131218): F: 5'-CACGGACATGATATCCATGCAGAC-3'; R: 5'-CTAGCTCACTTCTGACCATCAGATGC-3'; *Drbrn3c* (NM\_131278): F: 5'-ATGATGACATGAACGGCAAGC-3'; R: 5'-GTGCACTGCTGAATACTTCATCC-3'.

### Phylogenetic analysis of Dmbx proteins

Candidate Dmbx sequences were assembled as an ungapped Fasta file and imported into the Toffee server (<http://tcoffee.crg.cat>) and

processed using default parameters in an Expresso (<http://tcoffee.crg.cat/apps/tcoffee/do:expresso>) into a protein alignment<sup>27-31</sup>. The Toffee fasta alignment was imported into MegAlign Pro (DNASTAR ver 15.0.0) and ambiguous regions with poor alignment scores were removed, leaving only large, contiguous regions of well-aligned sequence. This alignment of 218 amino acid residues was exported in nexus format. The start of the *P. marinus* sequence is missing, and so residues were recoded from gaps to indicate missing sequence. The file was modified to include a MrBayes block, with aamodelpr = mixed, stopval = 0.01, ngen = 200000, and burninfrac = 0.25. The file was executed within MrBayes 3.2.1, and resulting consensus tree visualized in FigTree v.1.4.2 to show posterior probabilities (as percentage) at corresponding branch labels. Image output files from MegAlign Pro and FigTree v.1.4.2 were combined in Adobe Illustrator 2019 (Adobe Creative Suite 2019) (Extended Data Fig. 2). NCBI accession numbers or Ensembl identifiers for Dmbx sequences used in phylogenetic analyses are as follows: XP\_003725762.1 (*S. purpuratus*); NP\_001161526.1 (*S. kowalevskii*); AAT66431.1 (*B. floridae*); Ensembl ENSPMAG00000008114 (*P. marinus*); XP\_020369662.1 (*R. typos*); AAI34895.1 (*Dmbx1a*, *D. rerio*); NP\_001017625.1 (*Dmbx1b*, *D. rerio*); XP\_017949066.1 (*X. tropicalis*); XP\_001234036.2 (*G. gallus*); NP\_671725.1 (*H. sapiens*).

### In situ hybridization, sectioning, imaging, and biotapestry modelling

Whole mount in situ hybridization was performed using previously published protocols<sup>8,16</sup>. Cryosections of lamprey, skate, or zebrafish embryo in situ were sectioned at 18 µm with a Microm HM550 cryostat. In situ analysis of S25 skate embryos sections was performed using paraffin sections as follows: after fixation, embryos were embedded in paraffin and sections were prepared at 5-µm thickness on a Zeiss microtome. After paraffin removal with histosol, sections were hybridized with 1 ng/µl antisense digoxigenin-labelled probes overnight at 70 °C in a humidifying chamber. After hybridization, sections were washed with 50% formamide/50% 1× SSCT buffer followed by washes with MABT and a blocking step in 1% Roche blocking reagent. Sections were then incubated overnight at room temperature with a 1:2,000 dilution of anti-DIG-alkaline phosphatase antibody (Roche). After several washes with MABT, chromogenic colour was developed using NBT/BCIP precipitation (Roche). Imaging was performed on a Zeiss AxioImager.M2 equipped with an Apotome.2. Gene network models were assembled using Biotapestry<sup>32</sup>.

### Chicken embryo electroporation, dissociation, and cell sorting

Cranial and trunk NC cells were labelled using the previously published NC enhancers FoxD3-NC1.1 and FoxD3-NC2, respectively<sup>33</sup>. To isolate cranial NC cells, stage HH4 embryos were bilaterally electroporated with FoxD3-NC1.1>eGFP and cultured ex ovo until stage HH9+<sup>34</sup>. For each biological replicate, at least 15 embryo heads were dissected in Ringer's solution and washed three times in chilled 1× PBS. For trunk NC cells, stage HH10 embryos were bilaterally electroporated with FoxD3-NC2>eGFP and cultured in ovo until stage HH18. Five embryo trunks spanning the length of five somites were dissected in Ringer's solution and washed three times in chilled 1× PBS. The tissues were dissociated in Accumax (Innovative Cell Technologies, Inc.) for 15 min at 37 °C and GFP+ cells were collected using FACS.

### Library preparation and sequencing

Chicken libraries were prepared using the SMART-Seq Ultra Low Input RNA Kit (Takara) according to the manufacturer's protocol. For lamprey embryos, tissue was dissected from the cranial dorsal neural tubes of *n* = 100 T21 and trunk neural tubes of *n* = 100 T23.5 embryos. Total RNA was extracted using the RNAqueous kit (Ambion). RNA-seq was performed at the Millard and Muriel Jacobs Genetics and Genomics Laboratory (California Institute of Technology, Pasadena, CA) at 50 million, single-end reads on two biological replicates for both the



T21 cranial and T23.5 trunk neural tube samples. Sequencing libraries were built according to Illumina Standard Protocols. SR50 sequencing was performed on a HiSeq Illumina machine.

## Statistical analysis of lamprey and chicken axial population RNA-seq

To identify orthologous genes between lamprey and chicken, the lamprey proteome obtained from SIMRbase<sup>35</sup> was aligned to the chicken proteome using the BLAT alignment software available on the UCSC genome browser<sup>17,36</sup>. In brief, every lamprey protein sequence was queried locally against the chicken proteome, following which regions with the longest alignment were matched to the respective chicken genes. Using this alignment-based approach, proteins with an alignment percentage score between 52 and 100 (see Supplementary Table 1 for exact scores for each orthologue) were identified as orthologues, and their respective cDNA sequences were obtained from the chicken and lamprey databases. Chicken cranial and trunk libraries were aligned to the chicken sequences, while the lamprey cranial and trunk libraries were aligned to the lamprey sequences using Bowtie2<sup>37</sup>. Transcript counts were calculated using HTSeq-Count and differential gene expression analysis was performed using DESeq2<sup>38,39</sup>. Using chicken gene annotations as a reference, we added the transcript counts for duplicated orthologues found in the lamprey genome to calculate an 'aggregated' transcript count for each gene. These aggregated transcript counts were then normalized using the formula:

$$Z_i = \frac{T_i - \min(T)}{\max(T) - \min(T)}$$

where  $Z_i$  is the normalized transcript count and  $T_i$  is the absolute transcript count. A subset of genes previously identified as being part of the neural crest gene regulatory network<sup>15</sup> was then isolated from the count matrix and plotted as a heatmap to obtain the gene expression matrix.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

All raw sequencing data for all RNA-seq libraries (Fig. 3) and merged reference transcriptomes are available online (NCBI BioProject# PRJNA497902). Sequences of in situ probe templates for Figs. 1b, c, 2a, c are available through GenBank accession codes (see Methods).

## Code availability

Code used to analyse sequencing datasets are available from the corresponding author upon request.

26. Nikitina, N., Bronner-Fraser, M. & Sauka-Spengler, T. Culturing lamprey embryos. *Cold Spring Harbor Protoc.* <https://doi.org/10.1101/pdb.prot5122> (2009).
27. Di Tommaso, P. et al. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* **39**, W13–W17 (2011).
28. Armougom, F. et al. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.* **34**, W604–W608 (2006).
29. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G. & Notredame, C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.* **340**, 385–395 (2004).
30. Poirot, O., Suhre, K., Abergel, C., O'Toole, E. & Notredame, C. 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Res.* **32**, W37–W40 (2004).
31. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
32. Longabaugh, W. J. R., Davidson, E. H. & Bolouri, H. Computational representation of developmental genetic regulatory networks. *Dev. Biol.* **283**, 1–16 (2005).
33. Simões-Costa, M. S., McKeown, S. J., Tan-Cabugao, J., Sauka-Spengler, T. & Bronner, M. E. Dynamic and differential regulation of stem cell factor FoxD3 in the neural crest is encrypted in the genome. *PLoS Genet.* **8**, e1003142 (2012).
34. Hamburger, V. & Hamilton, H. L. A series of normal stages in the development of the chick embryo. *J. Morphol.* **88**, 49–92 (1951).
35. Smith, J. J. et al. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nat. Genet.* **50**, 270–277 (2018).
36. Karolchik, D. et al. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
37. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
38. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
39. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
40. Damas, H. Recherches sur le développement de *Lampetra fluviatilis* L. Contribution à l'étude de la céphalogenèse des vertébrés. *Arch. Biol. (Liege)* **1**–284 (1944).
41. Modrell, M. S. et al. A fate-map for cranial sensory ganglia in the sea lamprey. *Dev. Biol.* **385**, 405–416 (2014).

**Acknowledgements** We thank J. Tan-Cabugao and E. Grossman for technical assistance; D. Mayorga and R. Fraser for help with fish husbandry; B. Martik for illustrating the adult animals for our expression matrices; the Caltech Millard and Muriel Jacobs Genetics and Genomics Laboratory and in particular I. Antoshechkin for sequencing of our RNA-seq libraries; and R. Diamond, J. Tijerina, D. Perez, and P. Cannon of the The Caltech Flow Cytometry Cell Sorting Facility for cell sorting assistance. This work is supported by NIH grants R01NS086907, R01DE024157, and R35NS111564 to M.E.B. M.L.M. is supported by a Helen Hay Whitney Foundation postdoctoral fellowship. S.G. is supported by a graduate fellowship from the American Heart Association (18PRE34050063).

**Author contributions** Project and analysis conception were designed by M.L.M., M.S.-C., and M.E.B. Writing and interpretation were performed by M.L.M., S.G., B.R.U., J.A.G., S.A.G., M.S.-C., and M.E.B. Lamprey orthologue cloning and all in situ hybridization, imaging, and analysis were performed by M.L.M. Bioinformatics and chicken RNA-seq were performed by S.G. Phylogenetic analysis and lamprey embryo acquisition were performed by S.A.G. Cloning of skate orthologues and skate embryo acquisition were performed by J.A.G. Lamprey embryo dissections and library preparations were performed by B.R.U. and M.S.C.

**Competing interests** The authors declare no competing interests.

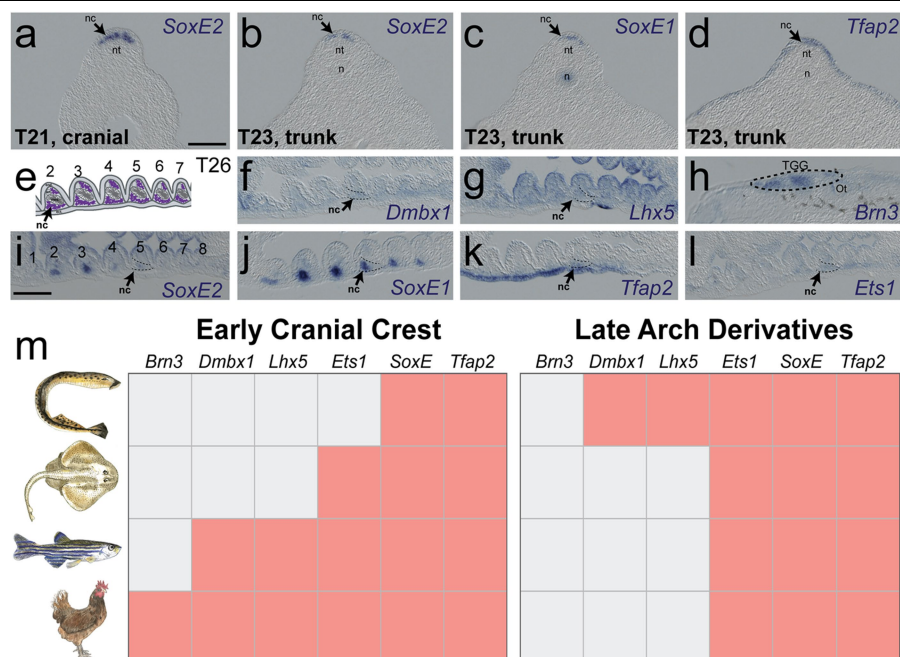
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1691-4>.

**Correspondence and requests for materials** should be addressed to M.E.B.

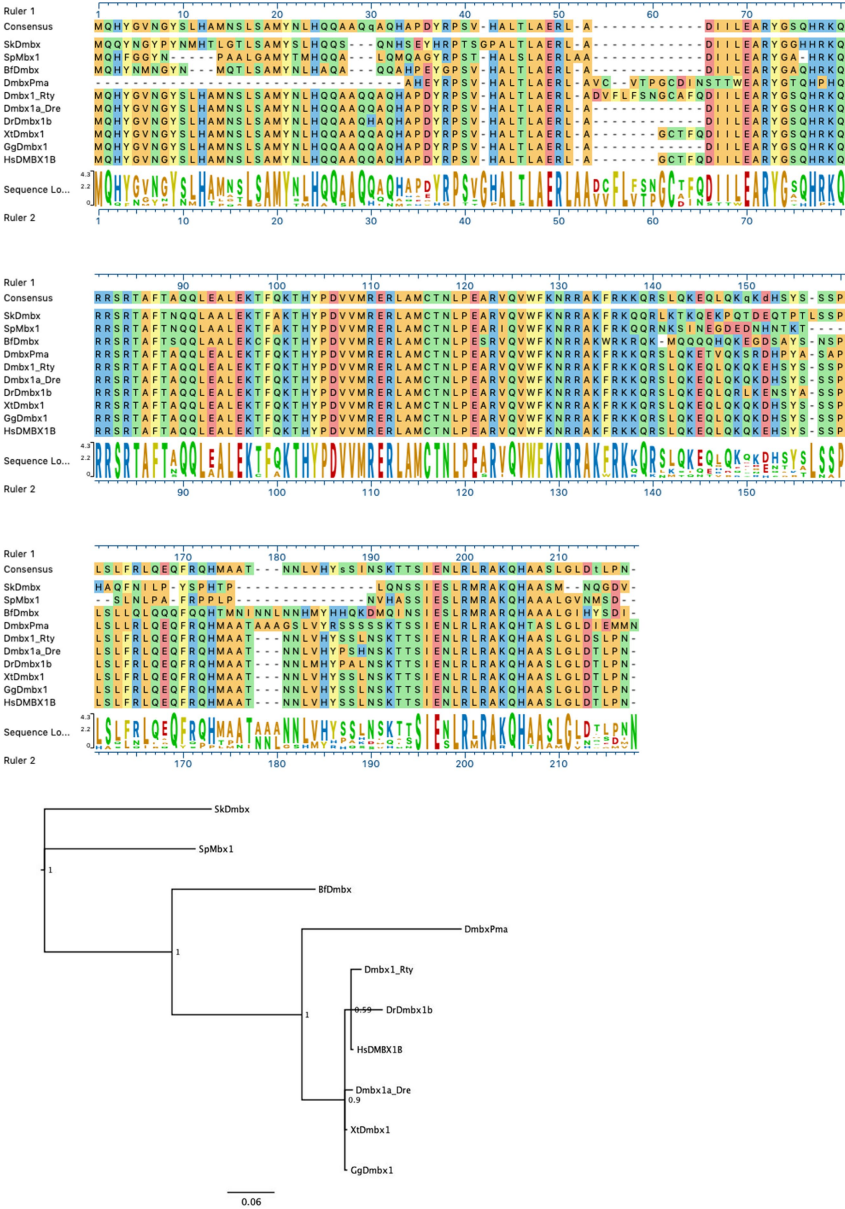
**Peer review information** Nature thanks Robert Cerny and Jeremiah Smith for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



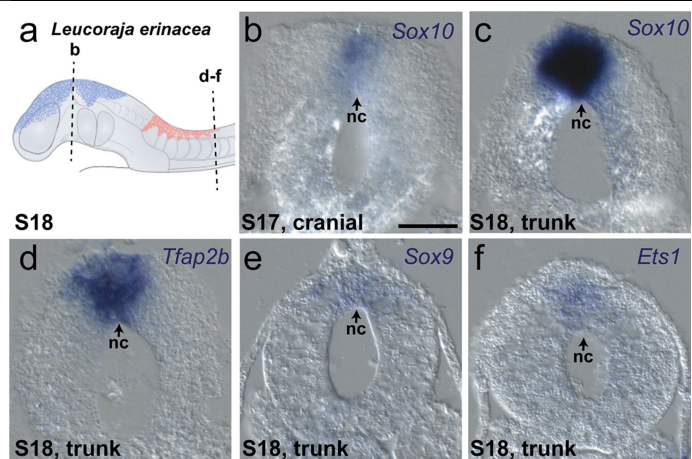
**Extended Data Fig. 1 | Heterochronic shifts of cranial specific gene regulatory nodes from later neural crest derivatives to an early specification program throughout gnathostome evolution.** **a–d**, Expression of lamprey orthologues of amniote cranial NC-specific genes at T21 (cranial) and T23 (trunk) in cross-section. **e**, Pharyngeal NC derivative expression in T26 *P. marinus* frontal section (based on ref. 40). **f–l**, Cranial circuit orthologues are

expressed in pharyngeal arch derivatives, with the exception of *Brn3*, which is expressed in the NC-derived cranial sensory ganglia<sup>41</sup> in lamprey frontal sections. **m**, Gene expression matrix summarizing the heterochronic shift of cranial crest specific circuit nodes. nc, neural crest; nt, neural tube; n, notochord. Scale bars, 100  $\mu$ m. Cryosections of in situ hybridizations were reproducible on  $n \geq 5$  embryos per time point for  $n \geq 2$  experiments.



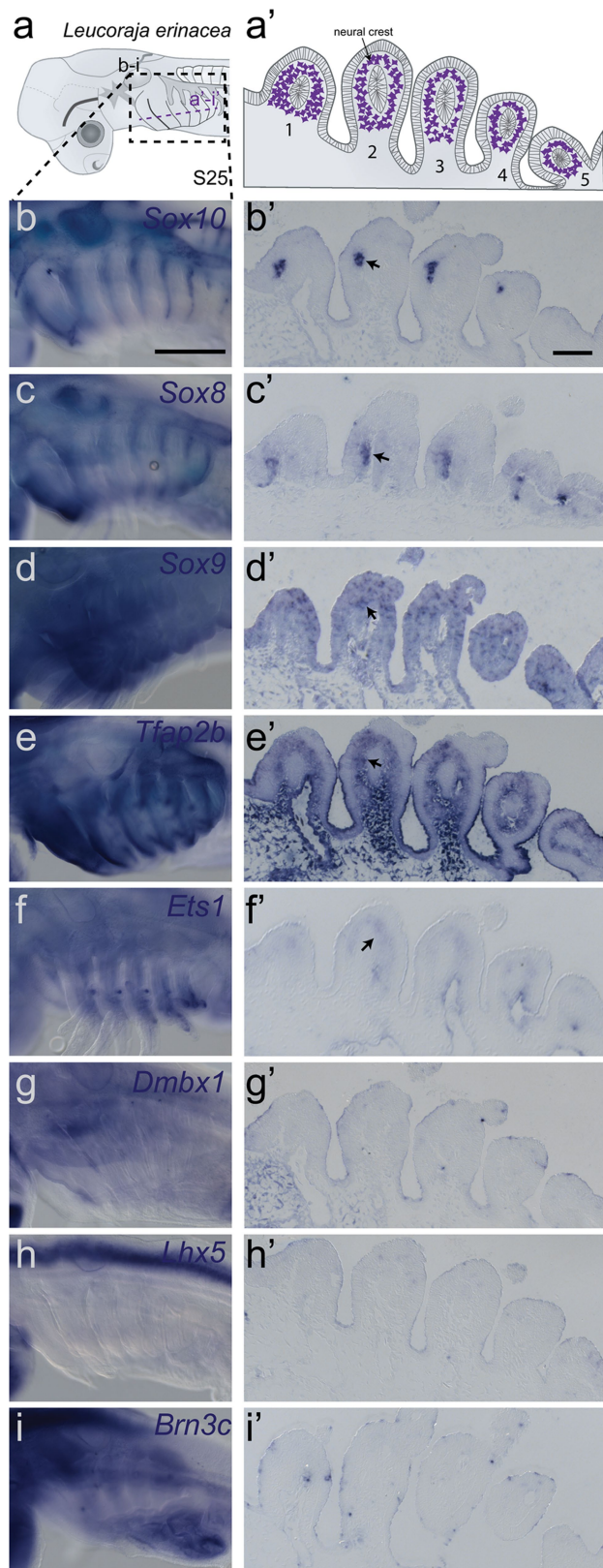
**Extended Data Fig. 2 | *P. marinus* Dmbx is homologous to gnathostome Dmbx genes. a**, Truncated alignment of Dmbx protein sequences. An alignment of full-length Dmbx protein sequences was assembled using T-Coffee and

contiguous regions tagged by the program as poorly or moderately well-aligned were removed, leaving 218 well-aligned residues. **b**, Bayesian consensus phylogenetic tree, with posterior probabilities shown at corresponding nodes.

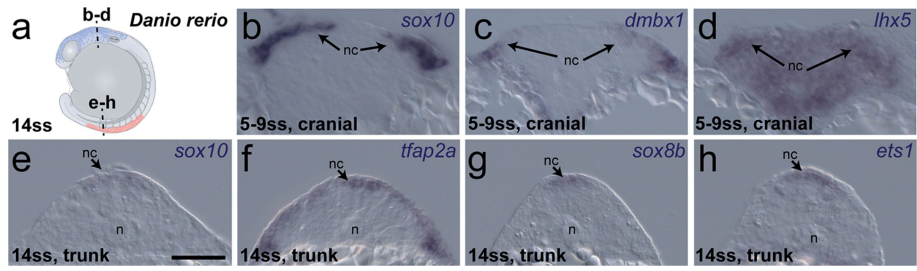


**Extended Data Fig. 3 | Expression of cranial circuit genes in the neural crest of the little skate.** **a**, Schematic of a stage 18 *L. erinacea* embryo with the NC illustrated as blue (cranial) and red (trunk). **b–f**, Cross-sections as depicted in **a**. Scale bar, 50  $\mu$ m. Cryosections of in situ hybridizations were reproducible on  $n \geq 2$  embryos for  $n \geq 2$  experiments.

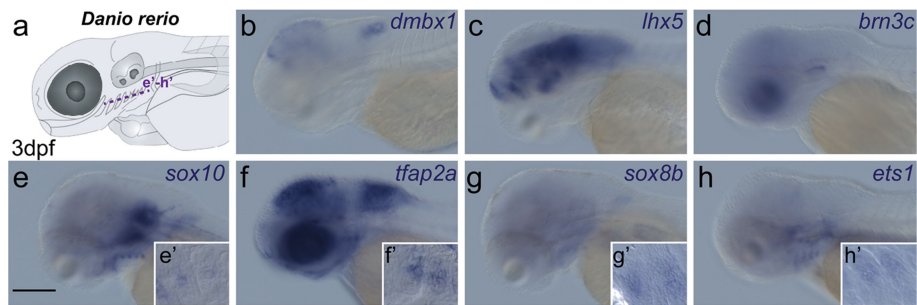




**Extended Data Fig. 4 | Pharyngeal neural crest derivative expression of cranial circuit orthologues in stage 25 *L. erinacea* embryos.** **a**, Schematic of head of *L. erinacea* embryo. Dashed box represents the region of the head for each embryo shown in **b–i** (left); purple dashed line depicts the location of the frontal section for **b–i** (right). **b–f**, Right, pharyngeal NC derivative expression of cranial circuit orthologues. **g–i**, Right, *Dmbx1*, *Lhx5*, and *Brn3c* are absent in pharyngeal arch derivatives at stage 25. **b–i**, Scale bars, 500  $\mu$ m (left); 100  $\mu$ m (right). In situ hybridizations were reproducible on  $n \geq 2$  embryos.



**Extended Data Fig. 5 | Expression of cranial circuit genes in the NC of the zebrafish.** **a**, Schematic of a 14ss *D. rerio* embryo with the NC illustrated as blue (cranial) and red (trunk). **b–h**, Cross-sections as shown in **a**. Scale bars, 50 μm. In situ hybridizations were reproducible on  $n \geq 10$  embryos.



**Extended Data Fig. 6 | Expression of cranial circuit orthologues in pharyngeal NC derivatives of *D. rerio* embryos at 3 days post fertilization.**  
**a**, Schematic of head of zebrafish embryo at 3 days post fertilization (dpf). Purple dashed line depicts the location of the frontal sections for the insets

in **e–h**. **e–h**, Expression of cranial circuit orthologues in pharyngeal arches. **b–d**, *Dmbx1*, *Lhx5*, and *Brn3c* are absent from pharyngeal arch derivatives at 3 dpf. Scale bar, 150  $\mu$ m. In situ hybridizations were reproducible on  $n \geq 10$  whole-mount and cryosectioned embryos.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☒ ☐ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated
- ☒ ☐ Clearly defined error bars  
*State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on [statistics for biologists](#) may be useful.

### Software and code

Policy information about [availability of computer code](#)

Data collection Zeiss Zen, Illumina HiSeq2500, WinList from Verity Software House

Data analysis BLAT alignment software, R v3.6.1, Bowtie2, HTSeq-Count, DESeq2, TCCoffee, Espresso, MegAlign Pro (DNASTAR 15.0.0), MrBayes3.2.1, FigTree 1.4.2, Biotapestry, Adobe Illustrator CS5.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All raw sequencing data for all RNAseq libraries (Figure 3) and merged reference transcriptomes are available online (NCBI BioProject# PRJNA497902). Sequences of in situ probe templates for Figures 1B, 1C, 2A, and 2C are available through GenBank accession codes found in the methods.



## Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For chicken RNAseq replicates, a pre-determined number of a cells was collected per kit instructions for the SMART-seq v4 ultra low input RNA kit (Cranial Rep 1= 1534 cells, Cranial Rep 2= 1530 cells, Cranial Rep 3= 1527 cells; Trunk Rep 1= 1500, Trunk Rep 2= 721, Trunk Rep 3= 958). For lamprey dissections, approximately 100 embryos were dissected for each replicate of each axial level for RNAseq libraries. No statistical tests were used to determine sample size. The sample size provided enough RNA for subsequent library preparation.
Data exclusions	No datasets were excluded from this analysis.
Replication	Chicken RNAseq libraries were collected in triplicate for biological replicates. Lamprey RNAseq libraries were collected with replicates, as well, for biological replicates. All attempts at replication were successful. For in situ hybridization, embryos were pooled from different breeding pairs (fish), brooding stocks (skates), or embryo batches (lamprey) to ensure replication of results in multiple fixed collections. All in situ expression patterns were replicated 100% over multiple rounds of in situs.
Randomization	Chicken eggs obtained from a local chicken farm were incubated in multiple incubators over different days to account for inter-batch variability. Lamprey tissues were collected from different breeding events to account for inter-batch variability.
Blinding	Different batches of chicken embryos were separately incubated in different incubators and electroporated with fresh DNA reporter construct solution. Different breeding pairs for lamprey were used for tissue collections to blind for batch variability. For all trunk libraries, dissections were made at the same somitic levels and not based on reporter expression to ensure blinding in sample collection.

## Reporting for specific materials, systems and methods

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

### Antibodies

Antibodies used	anti-Digoxigenin-Alkaline Phosphatase Fab Fragments (Roche ref#11093274910)
Validation	From the supplier: The polyclonal antibody from sheep is specific to digoxigenin and digoxin and shows no cross-reactivity with other steroids, such as human estrogens and androgens. Concentrations of antibody used for each animal has been previously reported and validated as cited in the methods section.

### Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Adult (3 months - 2 years of age) Danio rerio, zebrafish, were maintained in the Beckman Institute Fish Facility at Caltech, and all animal and embryo work was compliant under approved IACUC protocol 1346. Adult zebrafish are bred via natural spawning, which is triggered by light-dark cycle. The fish are kept in a specialized recirculating aquatic system with mechanical, chemical, and biological filters to maintain water quality. The water is kept at 28 degrees C and highly oxygenated.
Wild animals	Gravid male and female sea lamprey (Petromyzon marinus) were caught in the wild and provided by the Great Lakes Fishery

## Wild animals

Commission, in cooperation with its partners at the USGS Hammond Bay Biological Station, USFWS Marquette Biological Station, and the Department of Fisheries and Oceans, Canada. They were sent overnight in chilled, oxygenated water to the lamprey facility at the California Institute of Technology, where they were maintained under the parameters set in accordance with the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health, with protocols approved by the Institutional Animal Care and Use Committees of the California Institute of Technology (lamprey, Protocol #1436-17). After spawning, the captive adult lamprey died of natural causes.

## Field-collected samples

Lamprey embryos were produced by in vitro fertilization at the California Institute of Technology lamprey facility, using captive gravid lamprey (*Petromyzon marinus*) provided by the Great Lakes Fishery Commission, in cooperation with its partners at the USGS Hammond Bay Biological Station, USFWS Marquette Biological Station, and the Department of Fisheries and Oceans, Canada. Lamprey were maintained with a water temperature of 10-18 degrees C on a 15:9h light:dark cycle.

All skate embryos were collected from wild-caught brood stock housed at the Marine Resources Centre of the Marine Biological Laboratory (MBL) in Woods Hole. Eggs were maintained in a flow-through seawater system with a water temp of 15 degrees C, on a 12h:12h light:dark cycle. Prior to fixation, all embryos were euthanized with an overdose of buffered MS-222 (1g/L in seawater). All embryos collection was performed in accordance with protocols approved by the MBL Institutional Animal Care and Use committee.

Fertilized chicken eggs were obtained from a local farm in Sylmar, CA. Developing chicken embryos were maintained at a temperature of 37 degrees C.

## Flow Cytometry

### Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

#### Sample preparation

Chicken embryos were dissected in Ringers and washed thrice in chilled 1x PBS. The tissues were dissociated in Accumax (Innovative Cell Technologies, Inc.) for 15 minutes at 37°C.

#### Instrument

Sony SY3200 Cell Sorter

#### Software

WinList from Verity Software House

#### Cell population abundance

Collected cells were obtained and analyzed on a hemocytometer for fluorescence and viability to ensure 100% purity.

#### Gating strategy

Gating was assigned according to standard protocols. Clear differentials between GFP+ and GFP- populations were observed.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

# One thousand plant transcriptomes and the phylogenomics of green plants

<https://doi.org/10.1038/s41586-019-1693-2>

Received: 17 November 2017

Accepted: 12 September 2019

Published online: 23 October 2019

Open access

One Thousand Plant Transcriptomes Initiative

Green plants (Viridiplantae) include around 450,000–500,000 species<sup>1,2</sup> of great diversity and have important roles in terrestrial and aquatic ecosystems. Here, as part of the One Thousand Plant Transcriptomes Initiative, we sequenced the vegetative transcriptomes of 1,124 species that span the diversity of plants in a broad sense (Archaeplastida), including green plants (Viridiplantae), glaucophytes (Glaucophyta) and red algae (Rhodophyta). Our analysis provides a robust phylogenomic framework for examining the evolution of green plants. Most inferred species relationships are well supported across multiple species tree and supermatrix analyses, but discordance among plastid and nuclear gene trees at a few important nodes highlights the complexity of plant genome evolution, including polyploidy, periods of rapid speciation, and extinction. Incomplete sorting of ancestral variation, polyploidization and massive expansions of gene families punctuate the evolutionary history of green plants. Notably, we find that large expansions of gene families preceded the origins of green plants, land plants and vascular plants, whereas whole-genome duplications are inferred to have occurred repeatedly throughout the evolution of flowering plants and ferns. The increasing availability of high-quality plant genome sequences and advances in functional genomics are enabling research on genome evolution across the green tree of life.

Viridiplantae comprise an estimated 450,000–500,000 species<sup>1,2</sup>, encompass a high level of diversity and evolutionary timescales<sup>3</sup>, and have important roles in all terrestrial and most aquatic ecosystems. This ecological diversity derives from developmental, morphological and physiological innovations that enabled the colonization and exploitation of novel and emergent habitats. These innovations include multicellularity and the development of the plant cuticle, protected embryos, stomata, vascular tissue, roots, ovules and seeds, and flowers and fruit (Fig. 1). Thus, plant evolution ultimately influenced environments globally and created a cascade of diversity in other lineages that span the tree of life. Plant diversity has also fuelled agricultural innovations and growth in the human population<sup>4</sup>.

Phylogenomic approaches are now widely used to resolve species relationships<sup>5</sup> as well as the evolution of genomes, gene families and gene function<sup>6</sup>. We used mostly vegetative transcriptomes for a broad taxonomic sampling of 1,124 species together with 31 published genomes to infer species relationships and characterize the relative timing of organ-ismal, molecular and functional diversification across green plants.

We evaluated gene history discordance among single-copy genes. This is expected in the face of rapid species diversification, owing to incomplete sorting of ancestral variation between speciation events<sup>7</sup>. Hybridization<sup>8</sup>, horizontal gene transfer<sup>9</sup>, gene loss following gene and genome duplications<sup>10</sup> and estimation error can also contribute to gene-tree discordance. Nevertheless, through rigorous gene and species tree analyses, we derived robust species tree estimates (Fig. 2 and Supplementary Figs. 1–3). Gene-family expansions and genome duplications are recognized sources of variation for the evolution of gene function

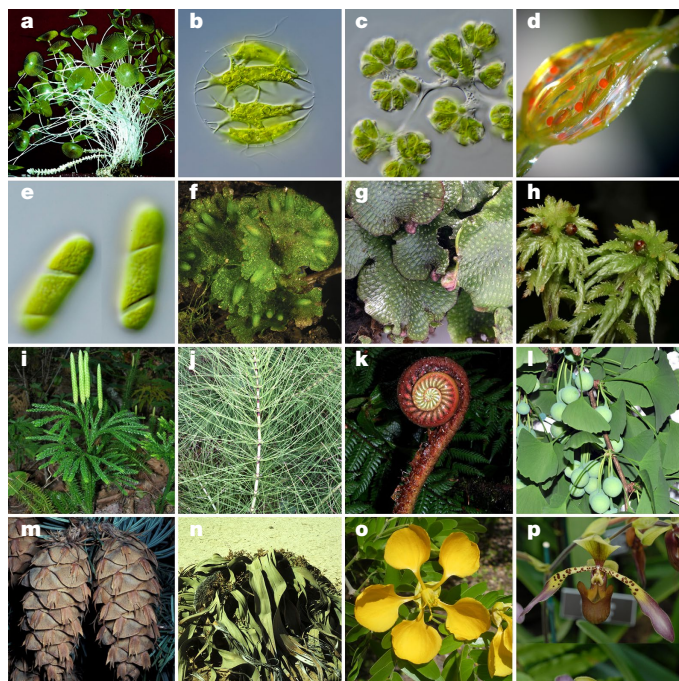
and biological innovations<sup>11,12</sup>. We inferred the timing of ancient genome duplications and large gene-family expansions. Our findings suggest that extensive gene-family expansions or genome duplications preceded the evolution of major innovations in the history of green plants.

## Integrated analysis of genome evolution

Because genome sizes vary by 2,340-fold in land plants<sup>13</sup> and 4,680-fold in chlorophyte and streptophyte green algae<sup>14</sup>, we used a reduced-representation sequencing approach to reconstruct gene and species histories. Specifically, we generated 1,342 transcriptomes representing 1,124 species across Archaeplastida, including green plants, glaucophytes and red algae. Comparing phylogenetic inferences based on nuclear and plastid genes (Figs. 2, 3 and Supplementary Figs. 1–3), we obtained well-supported, largely congruent results across diverse datasets and analyses. Resolution of some relationships, however, was confounded by gene-tree discordance (Fig. 3), which is attributable to factors that include rapid diversification, reticulate evolution, gene duplication and loss, and estimation error.

Inferred whole-genome duplications (WGDs; that is, polyploidy) across the gene-tree summary phylogeny estimated using ASTRAL<sup>15</sup> were not uniformly distributed (Fig. 4, Supplementary Fig. 8 and Supplementary Table 2). Comparing distributions of gene duplication times for each species<sup>16</sup> (Supplementary Table 3) and orthologue divergence times<sup>17</sup> (Supplementary Table 4) with gene-tree analyses<sup>18</sup> (Supplementary Tables 5, 6), we inferred 244 ancient WGDs across Viridiplantae (Supplementary Fig. 8 and Supplementary Table 2). Although there

A list of participants and their affiliations appears in the online version of the paper.



**Fig. 1 | Diversity within the Viridiplantae.** **a–e**, Green algae. **a**, *Acetabularia* sp. (Ulvophyceae). **b**, *Stephanosphaera pluvialis* (Chlorophyceae). **c**, *Botryococcus* sp. (Trebouxiophyceae). **d**, *Chara* sp. (Charophyceae). **e**, *Spirotaenia* sp. (taxonomy under review) (Zygnematophyceae). **f–p**, Land plants. **f**, *Notothylas orbicularis* (Anthocerotophyta (hornwort)). **g**, *Conocephalum conicum* (Marchantiophyta (thalloid liverwort)). **h**, *Sphagnum* sp. (Bryophyta (moss)). **i**, *Dendrolycopodium obscurum* (Lycopodiophyta (club moss)). **j**, *Equisetum telmateia* (Polypodiopsida, Equisetidae (horsetail)). **k**, *Parablechnum schiedeanum* (Polypodiopsida, Polypodiidae (leptosporangiate fern)). **l**, *Ginkgo biloba* (Ginkgophyta). **m**, *Pseudotsuga menziesii* (Pinophyta (conifer)). **n**, *Welwitschia mirabilis* (Gnetophyta). **o**, *Bulnesia arborea* (Angiospermae, eudicot, rosid). **p**, *Paphiopedilum lowii* (Angiospermae, monocot, orchid). **a**, Photograph reproduced with permission of Thieme Verlag, Stuttgart<sup>66</sup>. **b–e**, Photographs courtesy of M. Melkonian. **f–j**, **l–n**, **p**, Photographs courtesy of D.W.S. **k**, Photograph courtesy of R. Moran. **o**, Photograph courtesy of W. Judd.

are limitations to the inference of WGD events using this approach, we found that comparisons of these results with 65 overlapping published genome-based WGD inferences revealed 6 false-negative results in our tree-based estimates and no false-positive results (Supplementary Table 2). Analyses based on whole-genome sequences are needed for further resolution of WGD events.

With the exception of most *Selaginella* species and some liverworts (Fig. 1g), our analyses implicated at least one ancient WGD in the ancestry of every land plant lineage. By contrast, most algal lineages showed no evidence of WGD. Notably, the predicted sister clade of land plants (Fig. 2), Zygnematophyceae (Fig. 1e), exhibited the highest density of WGDs among algal lineages (Fig. 4), although the apparent increase in WGD was largely restricted to the desmid clade (Desmidiaceae) within Zygnematophyceae.

Increased diversification rates did not precisely co-occur with WGDs on the phylogeny. WGDs are expected to contribute to the evolution of novel gene function<sup>11,12</sup>. For example, novel functions among duplicate MADS-box genes that arose through WGD have been linked to the origin of flowering plants<sup>19,20</sup> and core eudicots<sup>21</sup>, and functional diversification of gene families after WGD has contributed to the evolution of fruit colour in tomato species<sup>22,23</sup> and to nodule development within legumes<sup>22,24</sup>. Consistent with previous studies with less extensive taxon sampling<sup>24–27</sup>, however, we inferred lags between WGDs and increased species diversity. Integrated phylogenomic and functional investigations are required to gain a mechanistic understanding of the lag

between WGD, the evolution of novel gene functions and their potential influence on diversification rates.

Gene-family expansions (and contractions) contribute to the dynamic evolution of metabolic, regulatory and signalling networks<sup>28,29</sup>. Given the inherent limitations of transcriptome data, we searched for large-fold changes in 23 of the largest gene families in *Arabidopsis thaliana*<sup>30</sup> that are involved in many important functions (such as transcriptional regulation, enzymatic and signalling function, and transport; Fig. 5 and Supplementary Tables 7, 8). Although our RNA-sequencing-based sampling of expressed genes is incomplete, the median representation of universally conserved genes<sup>31</sup> was 80–90% for taxa across Viridiplantae (Extended Data Fig. 3a, b). Furthermore, there was a strong correlation ( $r = 0.95$ ) between gene-family sizes in our transcriptomes (focusing on the largest gene families) and those of fully sequenced genomes (Extended Data Fig. 3c–f). We identified gene-family expansions and contractions, including some that have been described previously<sup>32–34</sup>. Specifically, the *AP2*, *bHLH*, *bZip* and *WRKY* transcription factor families were inferred to be present in the last common ancestor of Viridiplantae, whereas the origin of *GRAS* and *NAC* genes occurred in early streptophytes after divergence from the chlorophyte algal lineage (Fig. 5). The highest concentration of expansion events was inferred along the ‘spine’ of the phylogeny between the origins of Viridiplantae and vascular plants (Fig. 5b and Supplementary Table 7). Expansions of some focal gene families also continued after the origin of embryophytes; however, no expansions occurred in association with the origin and radiation of angiosperms (Fig. 5). Gene-family expansions and functional diversification may have contributed to the adaptations required for life in terrestrial habitats, but the sizes of these focal gene families apparently stabilized in the face of continued gene duplication and loss throughout the evolution of vascular plants.

### Primary acquisition of the plastid

The primary acquisition of the plastid in an ancestor of extant Archaeplastida was a pivotal event in the history of life. All possible relationships among Viridiplantae, Glaucophyta and Rhodophyta have been hypothesized, with alternative implications for the gain and loss of characters<sup>35</sup> in the early history of the three lineages. Strong support for the sister relationship of Viridiplantae and Glaucophyta<sup>35</sup> (Figs. 2, 3a) found here indicates that ancestral red algae lost flagella and peptidoglycan biosynthesis, perhaps associated with a reduction in genome size<sup>36</sup>. Peptidoglycan biosynthesis was independently lost early in the evolution of Chlorophyta<sup>37</sup> and within angiosperms<sup>38</sup>.

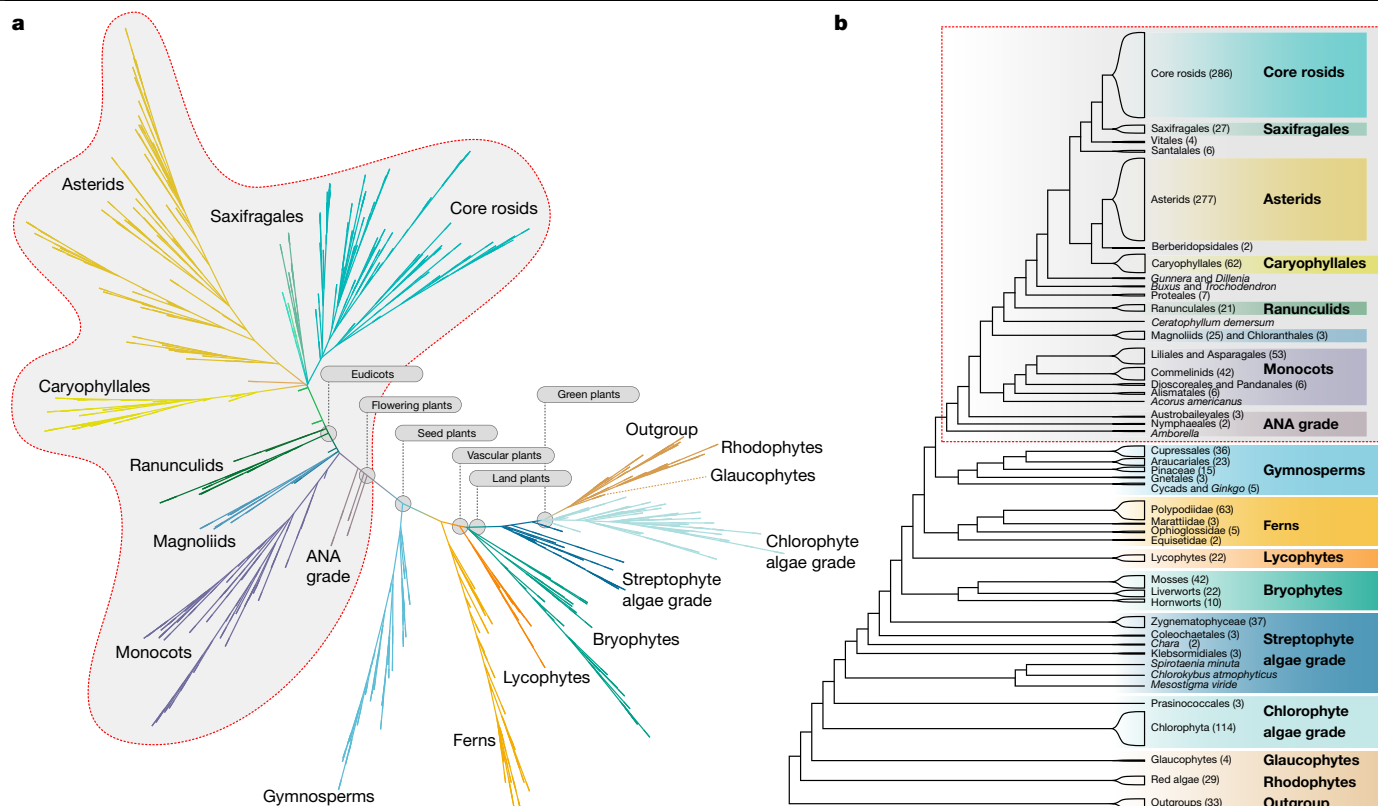
### The history of Viridiplantae

The origin of Viridiplantae is marked by the loss of light-harvesting phycobilisomes composed of phycobiliproteins, the evolution of the accessory photosynthetic pigment chlorophyll *b*, which has a distinct light-absorption spectrum relative to chlorophyll *a*, and intraplasmic starch synthesis and deposition. Viridiplantae are consistently recovered as monophyletic, with early diverging Chlorophyta and Streptophyta lineages<sup>39–41</sup>. However, the placement of the picoplanktonic algal lineage Prasinococcales was unstable in our analyses (Fig. 3e).

### Diversification within Chlorophyta

All nuclear-gene analyses resolved a grade of largely marine unicellular lineages subtending the core clade consisting of Trebouxiophyceae, Ulvophyceae and Chlorophyceae<sup>42</sup> (Fig. 1a–c and Supplementary Figs. 1–3). The nuclear supermatrix and ASTRAL trees placed Trebouxiophyceae as sister to a clade containing Chlorophyceae and Ulvophyceae<sup>42,43</sup>. However, whereas the supermatrix trees supported Ulvophyceae as monophyletic, the ASTRAL tree resolved Ulvophyceae as a grade and Bryopsidales is poorly supported as sister to Chlorophyceae (Fig. 3h). All tree estimates suggest that there were multiple origins of multicellularity





**Fig. 2 | Phylogenetic inferences of major clades.** Phylogenetic inferences were based on ASTRAL analysis of 410 single-copy nuclear gene families extracted from genome and transcriptome data from 1,153 species, including 1,090 green plant (Viridiplantae) species (Supplementary Table 1). **a**, Phylogram showing internal branch lengths proportional to coalescent units ( $2N_e$  generations) between branching events, as estimated by ASTRAL-II<sup>15</sup> v.5.0.3. **b**, Relationships

among major clades with red box outlining flowering plant clade. Species numbers are shown for each lineage. Most inferred relationships were robust across data types and analyses (Supplementary Figs. 1–3) with some exceptions (Supplementary Fig. 6). Data and analysis scripts are available at <https://doi.org/10.5281/zenodo.3255100>.

within Ulvophyceae. Only 12 out of 119 sampled chlorophyte species exhibited evidence of a WGD in their ancestry, and most of these putative WGDs were restricted to single clades.

### Streptophyta

The evolution of streptophytes was associated with several adaptations to terrestrial habitats<sup>44–46</sup>. All analyses recovered *Mesostigma*, *Chlorokybus* and *Spirotaenia minuta* in a clade that is sister to the remainder of Streptophyta<sup>39</sup> with successive divergence of Klebsormidiales, Charophyceae (Fig. 1d), Coleochaetophyceae and Zygnematophyceae (Fig. 1e) relative to Embryophyta. However, with greatly increased taxon sampling relative to our previous work<sup>39</sup>, internal branch lengths are diminished, and we could not reject the possibility of a true radiation giving rise to Coleochaetales, Zygnematophyceae and embryophyte lineages (land plants; Figs. 1f–p, 3g(II)). Although quartet support for a clade of Coleochaetales and Zygnematophyceae as sister to embryophytes was similar to support for Zygnematophyceae as sister to embryophytes, a clade consisting of Coleochaetales and land plants was not supported.

### Embryophyta

Land plants include many of the most familiar green plants (for example, bryophytes (Fig. 1f–h), lycophytes (Fig. 1i), ferns (Fig. 1j, k) and seed plants (Fig. 1l–p)). They exhibit key innovations, including protected reproductive organs (archegonia and antheridia) and the development of the zygote within an archegonium into an embryo that receives maternal nutrition. Resolving relationships among bryophytes (mosses, liverworts and hornworts) and their relationships to the remaining land plants has long been problematic, but is critical for understanding the

evolution of fundamental innovations within land plants, including the tolerance to desiccation, shifts in the dominance of multicellular haploid and diploid generations, and parental retention of a multicellular embryo.

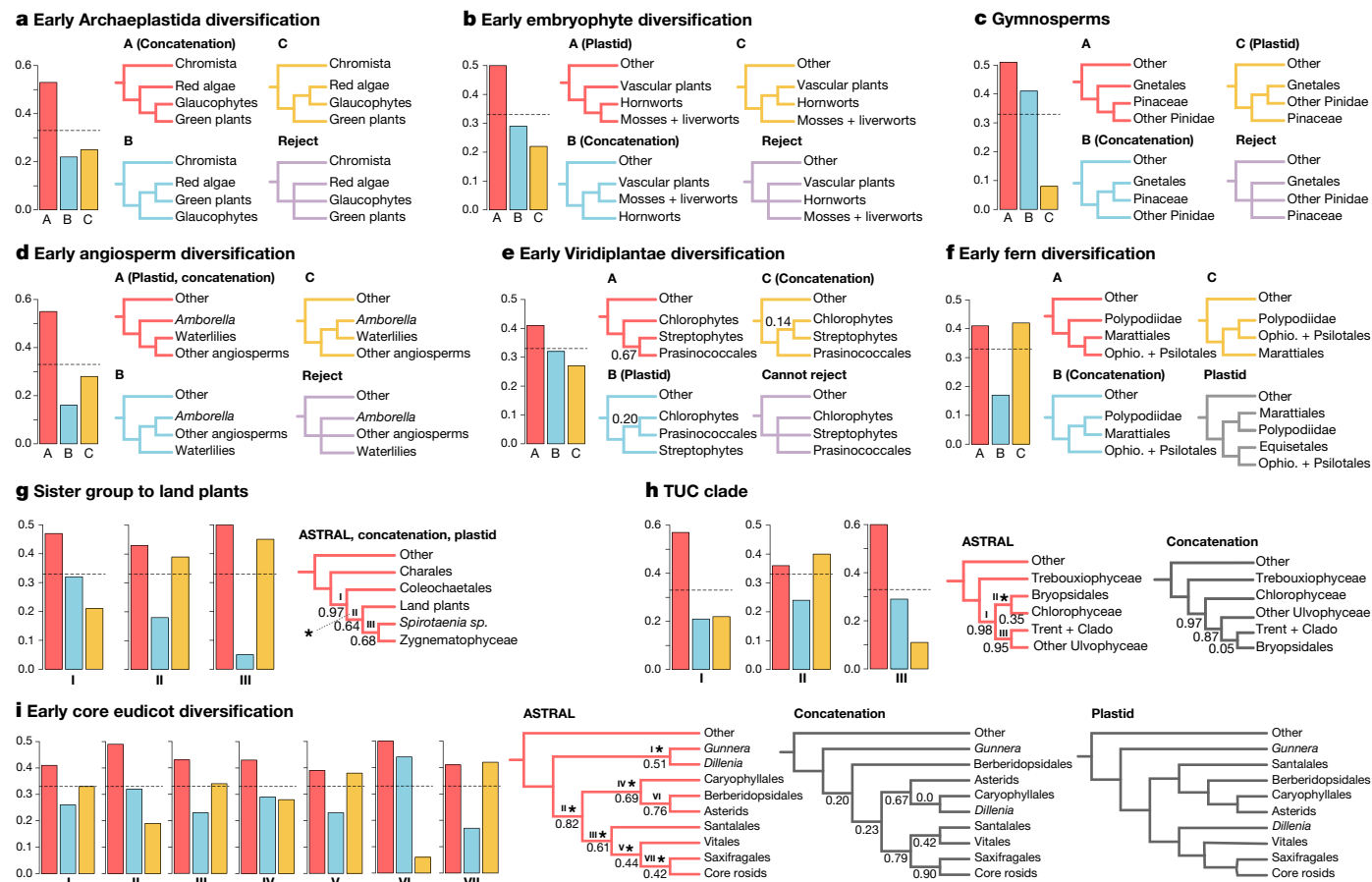
Bryophytes have sometimes been resolved as a grade<sup>47,48</sup>, with liverworts, mosses and hornworts as successive sister groups to Tracheophyta (vascular plants; Fig. 1i–p). We recovered extant bryophytes as monophyletic in the ASTRAL analysis of nuclear gene trees (Fig. 3b) and plastome analyses, with hornworts sister to a moss and liverwort clade. All analyses rejected the hypothesis that liverworts are sister to all other extant land plant lineages<sup>39,49</sup>.

The largest number of gene-family expansions in our analyses was associated with the origin of land plants and the evolution of bryophytes (transition between streptophyte algae and bryophytes in Fig. 5b). By contrast, we found no evidence of WGD on the main branch for land plants (Supplementary Tables 5, 6).

### Vascular plants

Within the vascular plants, lycophytes are supported as the sister group of Euphyllophyta (ferns and seed plants). We found no evidence of pan-vascular-plant or ancestral euphyllophyte WGDs, but some gene-family expansions were associated with the origin of vascular plants (Fig. 5b).

Within ferns (Polypodiopsida), plastid data weakly support Equisetales as sister to Psilotales and Ophioglossales (Supplementary Fig. 3), whereas nuclear gene analyses robustly place Equisetales sister to the remaining ferns<sup>50</sup>. The supermatrix and plastome-based trees placed Marattiiales sister to the leptosporangiate ferns<sup>50</sup> (Polypodiidae), but ASTRAL recovers nearly equal quartet support for this hypothesis or



**Fig. 3 | Alternative branching orders for contentious relationships.** Local posterior probabilities (shown only when below 1.0) and gene-tree quartet frequencies (bar graphs) for alternative branching orders for contentious relationships in the plant phylogeny (see text). **a**, Early Archaeplastida diversification. **b**, Early embryophyte diversification. **c**, Gymnosperms. **d**, Early angiosperm diversification. **e**, Early Viridiplantae diversification. **f**, Early fern diversification. **g**, The sister lineage to land plants. **h**, Trebouxiphyceae, Ulvophyceae and Chlorophyceae. **i**, Eudicot diversification. Red bars represent

the ASTRAL topology; blue and yellow trees and bars represent the frequencies of alternative branching orders in ASTRAL. The topologies recovered in the concatenated supermatrix analysis and plastid gene analyses are also indicated. Dashed horizontal lines mark expectation for a hard polytomy (purple). In **g–i**, panels include more than 4 tips, so nodes are delineated with Roman numerals and bar graphs are shown for each node and asterisks above branches indicate failure to reject the hypothesis that the node is a polytomy. Data and analysis scripts are available at <https://doi.org/10.5281/zenodo.3255100>.

for Marattiales as sister to Psilotales and Ophioglossales (Fig. 3f). Leptosporangiate ferns (Fig. 1k) experienced more WGD events than any other lineage of Viridiplantae outside the angiosperms, with an average of 3.79 inferred WGDs in the history of each sampled species (Fig. 4). WGD was inferred in an ancestor of all extant ferns and an additional 19 putative WGDs were implicated in the ancestry of fern subclades (Ophioglossaceae and Polypodiaceae; Fig. 4, Supplementary Fig. 8 and Supplementary Tables 2, 5, 6). Considering the high chromosome numbers of some ferns, our discovery that they exhibit one of the highest frequencies of palaeopolyploidization among green plants is not unexpected<sup>51</sup>.

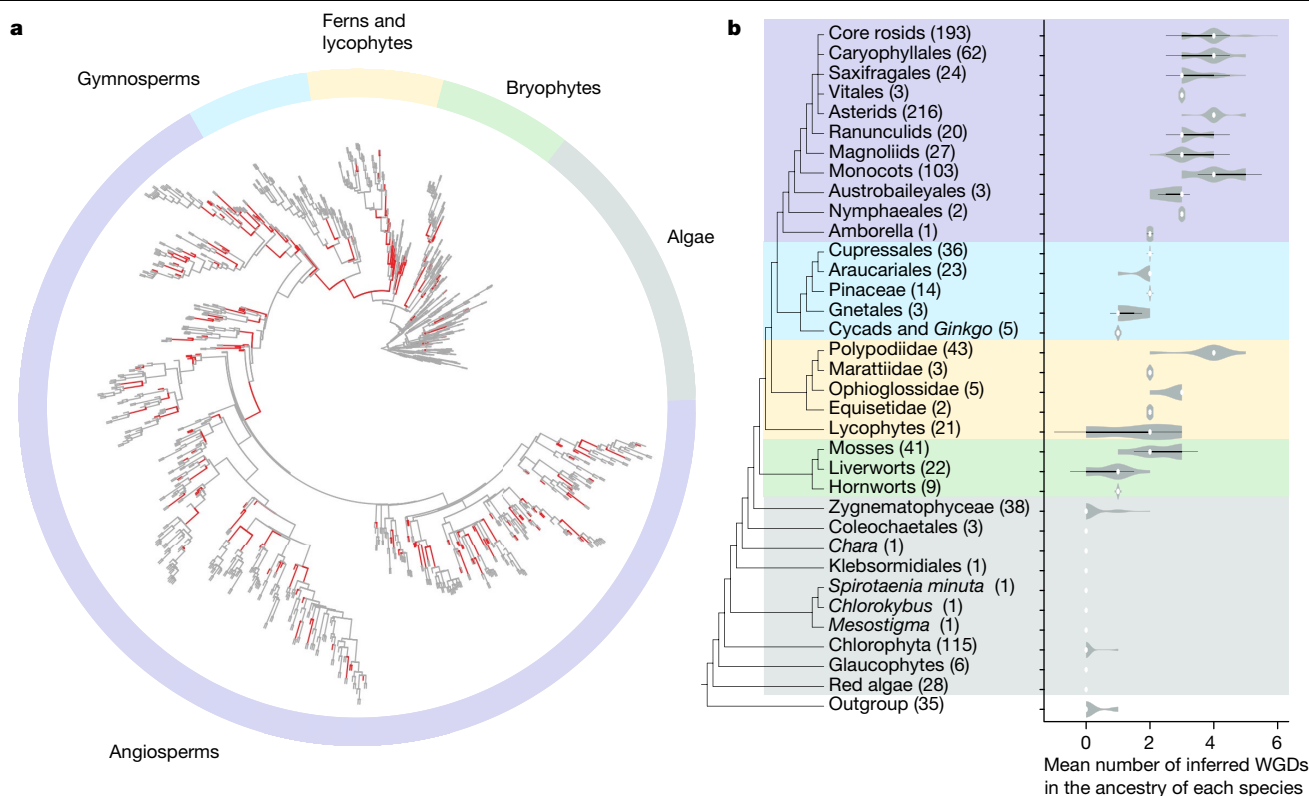
Whereas none of our focal gene families exhibited significant expansion in ferns, significantly more MIKC-type MADS-box genes—involved in specification of ovule and flower development in seed plants<sup>52</sup>—were observed in leptosporangiate ferns relative to all other green plant lineages, other than seed plants (Extended Data Fig. 1). The ancestral number of MIKC-type MADS-box genes for ferns and seed plants was 4 or 5, and gene numbers increased independently within leptosporangiate ferns and seed plants (Extended Data Figs. 1, 2).

### Seed plants

A WGD in the ancestry of all extant seed plants has been inferred previously<sup>18,53</sup> but remains contested<sup>54</sup>. Gene-tree<sup>18</sup> analyses revealed

significantly more gene duplications on the branch leading to extant seed plants than expected from background gene birth and death rates (analyses D1 ( $P < 2.0 \times 10^{-18}$ ) and D2 ( $P < 8.9 \times 10^{-16}$ ) in Supplementary Table 5). Numerous gene-family expansions were also associated with the origin of seed plants, and only one contraction was detected among the gene families analysed (Fig. 5b). Type II MIKC-type MADS-box genes exhibited a nearly twofold expansion independent of their expansion in ferns (Extended Data Figs. 1, 2).

Extant gymnosperms (approximately 1,000 species) are sister to flowering plants, and all of our analyses recovered Cycadales and *Ginkgo* (Fig. 1l) as a sister clade to the remaining gymnosperms (Fig. 3c). The placement of Gnetales conflicts strongly among the ASTRAL, supermatrix and plastome-based trees. Plastid data strongly support the ‘Gnecup’ hypothesis, with Gnetales as sister to a clade comprising Araucariales and Cupressales<sup>47</sup>, whereas the supermatrix analysis of nuclear genes supports a ‘Gnepine’ hypothesis with Gnetales as sister to Pinales<sup>55,56</sup>. ASTRAL analyses strongly support the ‘Gnetifer’ hypothesis, with conifers (Araucariales, Cupressales and Pinales) sister to Gnetales<sup>57</sup>. The short internal branches in the ASTRAL tree suggest rapid diversification (Fig. 2). However, the uneven frequencies of gene-tree quartets—which support the alternative Gnecup and Gnepine hypotheses—suggest that gene-tree estimation biases<sup>58</sup> associated with increased substitution rates in Gnetales<sup>59</sup> or gene flow are possible sources of gene-tree



**Fig. 4 | The distribution of inferred ancient WGDs across lineages of green plants. a**, The locations of estimated WGDs are labelled red in the phylogeny of all 1000 Plants (1KP) samples. **b**, The number of inferred ancient polyploidization events within each lineage is shown in the violin plots. The white dot indicates the median, the thick black bars represent the interquartile range, the thin black lines

define the 95% confidence interval and the grey shading represents the density of data points. The sample sizes for each lineage are shown within parentheses along with taxon names on the phylogeny. The phylogenetic placement of inferred WGDs is illustrated in Supplementary Fig. 8 and data supporting each WGD inference are provided in Supplementary Table 2.

discordance<sup>8</sup>. Previously inferred WGDs in ancestors of *Welwitschia*, Pinaceae and Cupressales<sup>18</sup> are supported, as is a new inference of WGD in the ancestry of Podocarpaceae (Fig. 4 and Supplementary Tables 2, 5, 6).

Angiosperms are by far the largest clade of green plants (more than 370,000 species<sup>2</sup>) and are marked by multiple key innovations, including the carpel, double fertilization, endosperm, and for most angiosperms, vessel elements. Both nuclear and plastid phylogenomic analyses agree with previous studies<sup>39</sup> in providing strong support for angiosperm monophyly and in placements of Amborellales, Nymphaeales and Austrobaileales as successive sisters to all other angiosperms (Figs. 2, 3). Chloranthales and magnoliids comprise a clade in the ASTRAL and supermatrix analyses, but were resolved with poor support as successive sister lineages to all other Mesangiospermae (monocots, *Ceratophyllum* and eudicots) in the plastome-based tree. Whereas *Ceratophyllum* is sister to eudicots in the ASTRAL and plastome trees, it is poorly supported as sister to monocots in the supermatrix tree (Supplementary Figs. 1–3). All analyses suggest short time intervals between branching of the monocots, Magnoliidae, Chloranthales, Ceratophyllales and eudicot lineages in early mesangiosperm history (Fig. 2 and Supplementary Figs. 1–3).

Pentapetalae (70% of all angiosperms) are marked by the evolution of the pentamerous flower. Substantial gene-tree discordance was observed for relationships among core rosids, Saxifragales, Vitales, *Dillenia*, Santalales, Berberidopsidales, Caryophyllales, asterids and Gunnerales (the sister group of Pentapetalae; Fig. 3i). Short internal branches and poor support in the ASTRAL tree at the base of the core eudicots (Figs. 2, 3i) indicate rapid diversification following two rounds of WGD that resulted in palaeohexaploidy preceding the origin of the clade<sup>60,61</sup> (Supplementary Fig. 8). The supermatrix and plastid trees conflict with the poorly supported ASTRAL branching order (Fig. 3i). With the exception of the Berberidopsidales and core asterid clade,

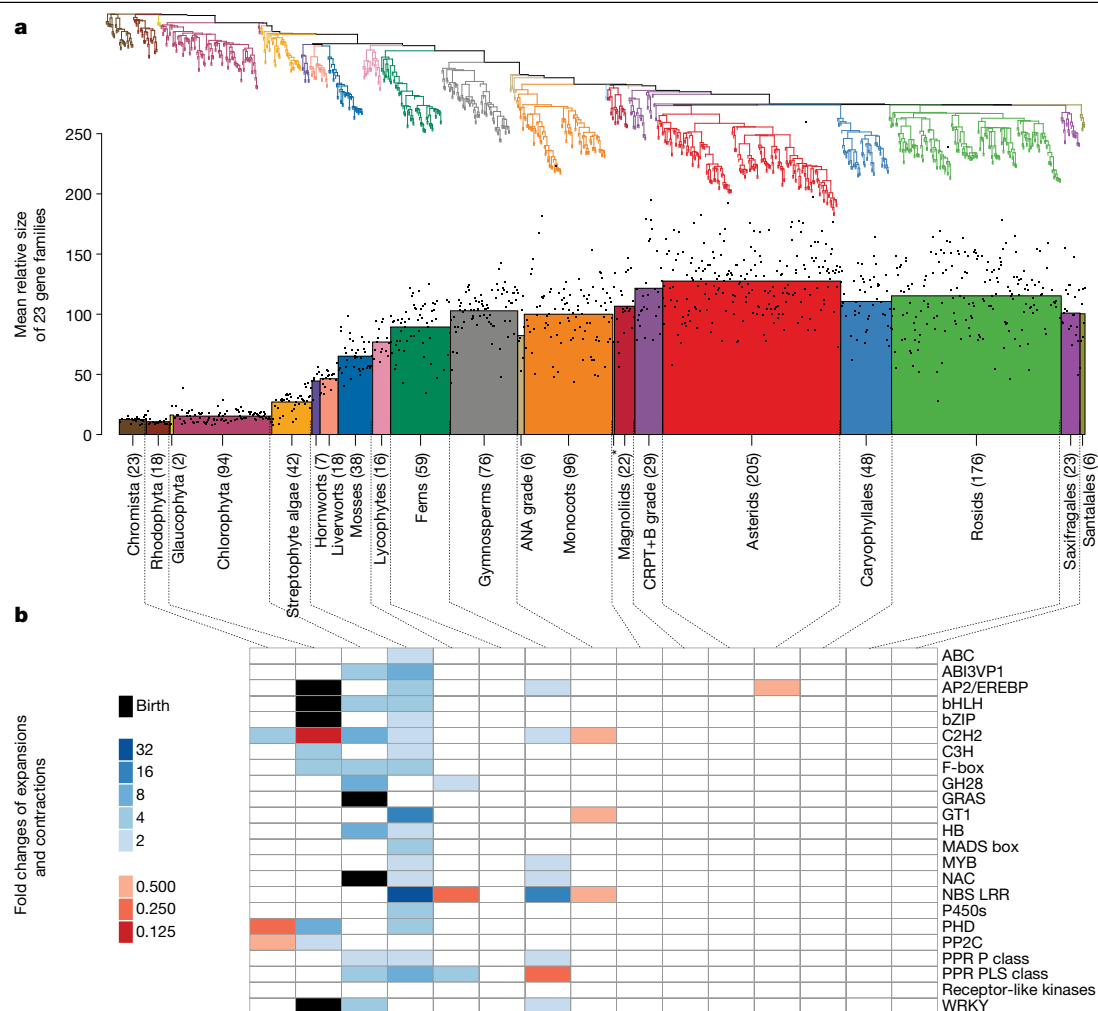
we were not able to reject the possibility of polytomies at the evaluated nodes in ASTRAL analyses (Fig. 3i).

Genomic and phylogenomic analyses have identified numerous WGDs throughout angiosperm history<sup>62,63</sup>. We found evidence that extant flowering plants descend from a polyploid common ancestor<sup>19,53</sup>. Gene-tree analyses detected a significantly larger-than-background number of gene duplications on the branch leading to the last common ancestor of extant angiosperms after divergence from the extant gymnosperm clade (analyses E1 ( $P < 1.8 \times 10^{-41}$ ) and E2 ( $1.4 \times 10^{-24}$ ) in Supplementary Table 5). Furthermore, the numbers of inferred duplications on the stem branch of angiosperms were consistent with expectations for WGD (analyses E1 and E2 in Supplementary Table 6). We inferred over 180 WGDs within flowering plants, including 132 in eudicots and 35 in monocots (Supplementary Table 2).

The origin of the angiosperms was preceded by three focal gene-family contractions and no expansions (Fig. 5b), consistent with the hypothesis that the innovations in angiosperms may have involved the functional co-option of genes that were duplicated earlier in the evolution of seed plants<sup>19</sup>. We find that orthologues of some floral homeotic MADS-box genes originated in the stem group of extant seed plants approximately 300 million years ago (Extended Data Fig. 2), supporting the hypothesis that the origin of the angiosperm flower involved recruitment of developmental regulators that already existed in their seed plant ancestors<sup>19,64</sup>.

## Synthesis

These analyses establish a foundation for advancing our understanding of the overall phylogenetic framework of green plants and the genetic changes that were responsible for the characteristic traits associated with major evolutionary transitions in Viridiplantae. Portions of the



**Fig. 5 | Assessment of significant expansions and contractions of largest plant gene families.** **a**, Weighted average gene-family size for species groups (normalized to account for differences in gene-family sizes, weight = 1/(maximum observed gene-family size)). The ANA grade comprises Amborellales, Nymphaeales and Austrobaileyales, successive sister lineages to a clade with the remaining extant angiosperms; the 'CRPT+B' grade includes Ceratophyllales, Ranunculales, Proteales lineages and a Trochodendrales + Buxales clade in the ASTRAL tree (Fig. 2). Sample sizes are proportional to bar widths (from left to right,  $n = 23$  (Chromista), 18 (Rhodophyta), 2 (Glaucophyta), 94 (Chlorophyta),

42 (streptophyte algae), 7 (hornworts), 18 (liverworts), 38 (mosses), 16 (lycophytes), 59 (ferns; monilophytes), 76 (gymnosperms), 6 (ANA grade), 96 (monocots), 1 (\*representing Chloranthales), 22 (magnoliids), 29 (CRPT+B grade), 205 (asterids), 48 (caryophyllids), 176 (rosids), 23 (Saxifragales) and 6 (Santalales). **b**, Gene families exhibiting significant copy number changes (two-sided Kolmogorov–Smirnov test;  $P < 1 \times 10^{-6}$ ; gene-family expansions represent a gain of more than 50% and contractions represent a loss of more than 33%) with colour codes showing the magnitude of the observed fold changes. Data and analysis scripts are available at <https://github.com/GrosseLab/OneKP-gene-family-evo>.

species tree reported here remain unresolved. Phylogenetic analyses of genes extracted from a broad sampling of whole-genome sequences may improve gene family circumscriptions and resolve the species tree further. Expanded genome sequencing may also help to accurately account for interspecific gene flow, and orthology in the face of gene duplications and losses. However, for some nodes in the species tree, extensive discordance among inferred gene histories suggests that rapid diversification may not always conform to strict bifurcation of ancestral species into two descendent species.

Gene and genome duplications have long been considered a source of evolutionary novelty<sup>11,12</sup>, producing an expanded molecular repertoire for adaptive evolution of key pathways and shifts in plant development and ecology. However, the direct connections between key innovations and specific gene duplications are rarely known, due in part to lag times between duplications and such innovations<sup>25–27</sup>. Phylogenetically informed experimental investigations of changes in gene content and function will improve our understanding of the roles of gene and genome duplications in the evolution of key innovations. Such efforts are underway, drawing on an expanding number

of experimental model species distributed across the green plant tree of life<sup>65</sup>.

## Online content


Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1693-2>

1. Corlett, R. T. Plant diversity in a changing world: status, trends, and conservation needs. *Plant Divers.* **38**, 10–16 (2016).
2. Lughadha, E. N. et al. Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. *Phytotaxa* **272**, 82–88 (2016).
3. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
4. Schery, R. W. *Plants for Man* 2nd edn (Prentice-Hall, 1972).
5. Philippe, H., Delsuc, F., Brinkmann, H. & Lartillot, N. Phylogenomics. *Annu. Rev. Ecol. Syst.* **36**, 541–562 (2005).



6. Eisen, J. A. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* **8**, 163–167 (1998).
7. Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340 (2009).
8. Solís-Lemus, C., Yang, M. & Ané, C. Inconsistency of species tree methods under gene flow. *Syst. Biol.* **65**, 843–851 (2016).
9. Yang, Z. et al. Horizontal gene transfer is more frequent with increased heterotrophy and contributes to parasite adaptation. *Proc. Natl Acad. Sci. USA* **113**, E7010–E7019 (2016).
10. Rasmussen, M. D. & Kellis, M. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* **22**, 755–765 (2012).
11. Ohno, S. *Evolution by Gene Duplication* (Springer-Verlag, 1970).
12. Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
13. Leitch, I. J. & Leitch, A. R. in *Plant Genome Diversity* Vol. 2 (eds Greilhuber, J. et al.) 307–322 (Springer, 2013).
14. Kapraun, D. F. Nuclear DNA content estimates in green algal lineages: chlorophyta and streptophyta. *Ann. Bot.* **99**, 677–701 (2007).
15. Mirarab, S. & Warnow, T. ASTRAL-III: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–i52 (2015).
16. Barker, M. S. et al. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* **25**, 2445–2455 (2008).
17. Barker, M. S. et al. EvoPipes.net: Bioinformatic Tools for Ecological and Evolutionary Genomics. *Evol. Bioinform. Online* **6**, 143–149 (2010).
18. Li, Z. et al. Early genome duplications in conifers and other seed plants. *Sci. Adv.* **1**, e1501084 (2015).
19. Amborella Genome Project. The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).
20. Ruelens, P. et al. The origin of floral organ identity quartets. *Plant Cell* **29**, 229–242 (2017).
21. Vekemans, D. et al. Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. *Mol. Biol. Evol.* **29**, 3793–3806 (2012).
22. Vanneste, K., Maere, S. & Van de Peer, Y. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Phil. Trans. R. Soc. B* **369**, 20130353 (2014).
23. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).
24. Cannon, S. B. et al. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Mol. Biol. Evol.* **32**, 193–210 (2015).
25. Schranz, M. E., Mohammadin, S. & Edger, P. P. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Curr. Opin. Plant Biol.* **15**, 147–153 (2012).
26. Tank, D. C. et al. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol.* **207**, 454–467 (2015).
27. Landis, J. B. et al. Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* **105**, 348–363 (2018).
28. Maere, S. et al. Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA* **102**, 5454–5459 (2005).
29. Hanada, K., Zou, C., Lehti-Shiu, M. D., Shinozaki, K. & Shiu, S.-H. Importance of lineage-specific expansion of plant tandem duplications in the adaptive response to environmental stimuli. *Plant Physiol.* **148**, 993–1003 (2008).
30. Nelson, D. & Werck-Reichhart, D. A P450-centric view of plant evolution. *Plant J.* **66**, 194–211 (2011).
31. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
32. Bowman, J. L. et al. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* **171**, 287–304 (2017).
33. Catarino, B., Hetherington, A. J., Emms, D. M., Kelly, S. & Dolan, L. The stepwise increase in the number of transcription factor families in the Precambrian predated the diversification of plants on land. *Mol. Biol. Evol.* **33**, 2815–2819 (2016).
34. Wilhelmsson, P. K. I., Mühlich, C., Ullrich, K. K. & Rensing, S. A. Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae. *Genome Biol. Evol.* **9**, 3384–3397 (2017).
35. Rodríguez-Ezpeleta, N. et al. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr. Biol.* **15**, 1325–1330 (2005).
36. Qiu, H., Price, D. C., Yang, E. C., Yoon, H. S. & Bhattacharya, D. Evidence of ancient genome reduction in red algae (Rhodophyta). *J. Phycol.* **51**, 624–636 (2015).
37. van Baren, M. J. et al. Evidence-based green algal genomics reveals marine diversity and ancestral characteristics of land plants. *BMC Genomics* **17**, 267 (2016).
38. Grosche, C. & Rensing, S. A. Three rings for the evolution of plastid shape: a tale of land plant FtsZ. *Protoplasma* **254**, 1879–1885 (2017).
39. Wickett, N. J. et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl Acad. Sci. USA* **111**, E4859–E4868 (2014).
40. Lewis, L. A. & McCourt, R. M. Green algae and the origin of land plants. *Am. J. Bot.* **91**, 1535–1556 (2004).
41. Becker, B. & Marin, B. Streptophyte algae and the origin of embryophytes. *Ann. Bot.* **103**, 999–1004 (2009).
42. Marin, B. Nested in the Chlorellales or independent class? Phylogeny and classification of the Pedinophyceae (Viridiplantae) revealed by molecular phylogenetic analyses of complete nuclear and plastid-encoded rRNA operons. *Protist* **163**, 778–805 (2012).
43. Cocquyt, E., Verbruggen, H., Leliaert, F. & De Clerck, O. Evolution and cytological diversification of the green seaweeds (Ulvoophyceae). *Mol. Biol. Evol.* **27**, 2052–2061 (2010).
44. Delaux, P.-M. et al. Algal ancestor of land plants was preadapted for symbiosis. *Proc. Natl Acad. Sci. USA* **112**, 13390–13395 (2015).
45. Maugarny-Calès, A. et al. Apparition of the NAC transcription factors predates the emergence of land plants. *Mol. Plant* **9**, 1345–1348 (2016).
46. Delwiche, C. F. & Cooper, E. D. The evolutionary origin of a terrestrial flora. *Curr. Biol.* **25**, R899–R910 (2015).
47. Nickrent, D. L., Parkinson, C. L., Palmer, J. D. & Duff, R. J. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol.* **17**, 1885–1895 (2000).
48. Shaw, A. J., Szövényi, P. & Shaw, B. Bryophyte diversity and evolution: windows into the early evolution of land plants. *Am. J. Bot.* **98**, 352–369 (2011).
49. Puttick, M. N. et al. The interrelationships of land plants and the nature of the ancestral embryophyte. *Curr. Biol.* **28**, 733–745 (2018).
50. Rothfels, C. J. et al. The evolutionary history of ferns inferred from 25 low-copy nuclear genes. *Am. J. Bot.* **102**, 1089–1107 (2015).
51. Barker, M. S. & Wolf, P. G. Unfurling fern biology in the genomics age. *Bioscience* **60**, 177–185 (2010).
52. Theißen, G. & Gramzow, L. in *Plant Transcription Factors: Evolutionary, Structural, and Functional Aspects* (ed. Gonzalez, D. H.) 127–138 (Academic, 2016).
53. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
54. Ruprecht, C. et al. Revisiting ancestral polyploidy in plants. *Sci. Adv.* **3**, e1603195 (2017).
55. Bowe, L. M., Coat, G. & dePamphilis, C. W. Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proc. Natl Acad. Sci. USA* **97**, 4092–4097 (2000).
56. Chaw, S. M., Parkinson, C. L., Cheng, Y., Vincent, T. M. & Palmer, J. D. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc. Natl Acad. Sci. USA* **97**, 4086–4091 (2000).
57. Chaw, S. M., Zharkikh, A., Sung, H. M., Lau, T. C. & Li, W. H. Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. *Mol. Biol. Evol.* **14**, 56–68 (1997).
58. Zhong, B., Yonezawa, T., Zhong, Y. & Hasegawa, M. The position of Gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. *Mol. Biol. Evol.* **27**, 2855–2863 (2010).
59. Wan, T. et al. A genome for gnetophytes and early evolution of seed plants. *Nat. Plants* **4**, 82–89 (2018).
60. Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
61. Jiao, Y. et al. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, R3 (2012).
62. Soltis, D. E. et al. Polyploidy and angiosperm diversification. *Am. J. Bot.* **96**, 336–348 (2009).
63. Vanneste, K., Baele, G., Maere, S. & Van de Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).
64. Moyroud, E. et al. A link between LEAFY and B-gene homologues in *Welwitschia mirabilis* sheds light on ancestral mechanisms prefiguring floral development. *New Phytol.* **216**, 469–481 (2017).
65. Chang, C., Bowman, J. L. & Meyerowitz, E. M. Field guide to plant model systems. *Cell* **167**, 325–339 (2016).
66. Berger, S. & Kaever, M. *J. Dasycladales: an Illustrated Monograph of a Fascinating Algal Order* (Thieme Verlag, 1992).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019



James H. Leebens-Mack<sup>1,135\*</sup>, Michael S. Barker<sup>2,135</sup>, Eric J. Carpenter<sup>3,134</sup>, Michael K. Deyholos<sup>4,135</sup>, Matthew A. Gitzendanner<sup>5,6,134</sup>, Sean W. Graham<sup>7,135</sup>, Ivo Grosse<sup>8,11,135</sup>, Zheng Li<sup>2,134</sup>, Michael Melkonian<sup>9,135</sup>, Siavash Mirarab<sup>10,134,135</sup>, Martin Porsch<sup>11,134</sup>, Marcel Quint<sup>12,135</sup>, Stefan A. Rensing<sup>13,14,135</sup>, Douglas E. Soltis<sup>5,15,135</sup>, Pamela S. Soltis<sup>5,15,135</sup>, Dennis W. Stevenson<sup>16,135</sup>, Kristian K. Ullrich<sup>17,134</sup>, Norman J. Wickett<sup>18,19</sup>, Lisa DeGironimo<sup>16,134</sup>, Patrick P. Edge<sup>20,134</sup>, Ingrid E. Jordan-Thaden<sup>5,6,21,134</sup>, Steve Joya<sup>7,134</sup>, Tao Liu<sup>22,134</sup>, Barbara Melkonian<sup>9,134</sup>, Nicholas W. Miles<sup>23,134</sup>, Lisa Pokorny<sup>24,25,26,134</sup>, Charlotte Quigley<sup>27,134</sup>, Philip Thomas<sup>28,134</sup>, Juan Carlos Villarreal<sup>29,134</sup>, Megan M. Augustin<sup>30</sup>, Matthew D. Barrett<sup>31,32,33</sup>, Regina S. Baucom<sup>34</sup>, David J. Beerling<sup>35</sup>, Ruben Maximilian Benstein<sup>36</sup>, Ed Biffin<sup>37</sup>, Samuel F. Brockington<sup>38</sup>, Dylan O. Burge<sup>39</sup>, Jason N. Burris<sup>40,41,42</sup>, Kellie P. Burris<sup>40,43</sup>, Valérie Burtet-Sarramegna<sup>44</sup>, Ana L. Caicedo<sup>45</sup>, Steven B. Cannon<sup>46</sup>, Zehra Çebi<sup>9</sup>, Ying Chang<sup>7,47</sup>, Caspar Chater<sup>48</sup>, John M. Cheeseman<sup>49</sup>, Tao Chen<sup>50</sup>, Neil D. Clarke<sup>51</sup>, Harmony Clayton<sup>52</sup>, Sarah Covichoff<sup>38</sup>, Barbara J. Crandall-Stotler<sup>53</sup>, Hugh Cross<sup>54</sup>, Claude W. dePamphilis<sup>55,134</sup>, Joshua P. Der<sup>56</sup>, Ron Determann<sup>57</sup>, Rowan C. Dickson<sup>58</sup>, Verónica S. Di Stilio<sup>39</sup>, Shona Ellis<sup>5</sup>, Eva Fast<sup>49</sup>, Nicole Feja<sup>9</sup>, Katie J. Field<sup>60</sup>, Dmitry A. Filatov<sup>61</sup>, Patrick M. Finnegan<sup>31</sup>, Sandra K. Floyd<sup>62</sup>, Bruno Fogliani<sup>44,63</sup>, Nicolás García<sup>64</sup>, Gildas Gâteblé<sup>69</sup>, Grant T. Godden<sup>6</sup>, Falicia (Qi Yun) Goh<sup>65</sup>, Stephan Greiner<sup>66</sup>, Alex Harkess<sup>130</sup>, James Mike Heaney<sup>5,6</sup>, Katherine E. Helliwell<sup>67,68</sup>, Karolina Heyduk<sup>1,69</sup>, Julian M. Hibberd<sup>38</sup>, Richard G. J. Hodel<sup>5,6,34</sup>, Peter M. Hollingsworth<sup>28</sup>, Marc T. J. Johnson<sup>70</sup>, Ricarda Jost<sup>31,71</sup>, Blake Joyce<sup>40,72</sup>, Maxim V. Kapralov<sup>73</sup>, Elena Kazamia<sup>38</sup>, Elizabeth A. Kellogg<sup>30,74</sup>, Marcus A. Koch<sup>75</sup>, Matt Von Konrat<sup>76</sup>, Kálmán Könyves<sup>77,78</sup>, Toni M. Kutchan<sup>30</sup>, Vivienne Lam<sup>7</sup>, Anders Larsson<sup>79</sup>, Andrew R. Leitch<sup>80</sup>, Roswitha Lentz<sup>2</sup>, Fay-Wei Li<sup>81</sup>, Andrew J. Lowe<sup>82</sup>, Martha Ludwig<sup>52</sup>, Paul S. Manos<sup>83</sup>, Evgeny Mavrodiev<sup>5,6</sup>, Melissa K. McCormick<sup>84</sup>, Michael McKain<sup>85</sup>, Tracy McLellan<sup>86</sup>, Joel R. McNeal<sup>1,87</sup>, Richard E. Miller<sup>88</sup>, Matthew N. Nelson<sup>89,90,91</sup>, Yanhui Peng<sup>40,92</sup>, Paula Ralph<sup>55</sup>, Daniel Real<sup>93</sup>, Chance W. Riggins<sup>94</sup>, Markus Ruhsam<sup>28</sup>, Rowan F. Sage<sup>95</sup>, Ann K. Sakai<sup>96</sup>, Moira Scascitella<sup>7</sup>, Edward E. Schilling<sup>97</sup>, Eva-Marie Schlösser<sup>98</sup>, Heike Sederoff<sup>98</sup>, Stein Servick<sup>5</sup>, Emily B. Sessa<sup>9</sup>, A. Jonathan Shaw<sup>83</sup>, Shane W. Shaw<sup>99</sup>, Erin M. Sigel<sup>100</sup>, Cynthia Skema<sup>101</sup>, Alison G. Smith<sup>38</sup>, Ann Smithson<sup>31</sup>, C. Neal Stewart Jr<sup>40,41</sup>, John R. Stinchcombe<sup>95,102</sup>, Peter Szövényi<sup>103</sup>, Jennifer A. Tate<sup>58</sup>, Helga Tiebel<sup>9</sup>, Dorset Trapnell<sup>1</sup>, Matthieu Villegente<sup>44</sup>, Chun-Neng Wang<sup>104</sup>, Stephen G. Weller<sup>96</sup>, Michael Wenzel<sup>57</sup>, Stina Weststrand<sup>105</sup>, James H. Westwood<sup>106</sup>, Dennis F. Whigham<sup>84</sup>, Shuangxiu Wu<sup>107,134</sup>, Adrien S. Wulff<sup>44,63</sup>, Yu Yang<sup>108</sup>, Dan Zhu<sup>109</sup>, Cuili Zhuang<sup>7</sup>, Jennifer Zuido<sup>110</sup>, Mark W. Chase<sup>26,111,135</sup>, J. Chris Pires<sup>112,134</sup>, Carl J. Rothfels<sup>83,113,114,134</sup>, Jun Yu<sup>107,134</sup>, Cui Chen<sup>115</sup>, Li Chen<sup>116</sup>, Shifeng Cheng<sup>117</sup>, Juanjuan Li<sup>116</sup>, Ran Li<sup>116</sup>, Xia Li<sup>116</sup>, Haorong Lu<sup>116</sup>, Yanxiang Ou<sup>116</sup>, Xiao Sun<sup>118</sup>, Xuemei Tan<sup>116</sup>, Jingbo Tang<sup>118</sup>, Zhijian Tian<sup>115</sup>, Feng Wang<sup>120</sup>, Jun Wang<sup>121</sup>, Xiaofeng Wei<sup>116</sup>, Xun Xu<sup>116</sup>, Zhixiang Yan<sup>116</sup>, Fan Yang<sup>116</sup>, Xiaoni Zhong<sup>118</sup>, Feiyu Zhou<sup>116</sup>, Ying Zhu<sup>116</sup>, Yong Zhang<sup>116,118,135</sup>, Saravanaraj Ayyampalayam<sup>1122</sup>, Todd J. Barkman<sup>123</sup>, Nam-phuong Nguyen<sup>124</sup>, Naim Matasiçi<sup>125</sup>, David R. Nelson<sup>126</sup>, Erfan Sayyari<sup>10</sup>, Eric K. Wafula<sup>55</sup>, Ramona L. Walls<sup>72</sup>, Tandy Warnow<sup>127,134</sup>, Hong An<sup>128</sup>, Nils Arrigo<sup>2</sup>, Anthony E. Baniaga<sup>2</sup>, Sally Galuska<sup>2</sup>, Stacy A. Jorgensen<sup>129</sup>, Thomas I. Kidder<sup>2</sup>, Hanghui Kong<sup>130</sup>, Patricia Lu-Irving<sup>2</sup>, Hannah E. Marx<sup>2,34</sup>, Xinchuai Qi<sup>2</sup>, Chris R. Reardon<sup>2</sup>, Brittany L. Sutherland<sup>2</sup>, George P. Tiley<sup>95</sup>, Shana R. Welles<sup>2</sup>, Rongpei Yu<sup>131</sup>, Shing Zhan<sup>113</sup>, Lydia Gramzow<sup>132</sup>, Günter Theißen<sup>132</sup> & Gane Ka-Shu Wong<sup>3,116,133,135\*</sup>

<sup>1</sup>Department of Plant Biology, University of Georgia, Athens, GA, USA. <sup>2</sup>Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA. <sup>3</sup>Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada. <sup>4</sup>Department of Biology, The University of British Columbia Okanagan, Kelowna, British Columbia, Canada. <sup>5</sup>Department of Biology, University of Florida, Gainesville, FL, USA. <sup>6</sup>Florida Museum of Natural History, University of Florida, Gainesville, FL, USA. <sup>7</sup>Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada. <sup>8</sup>German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Germany. <sup>9</sup>Botanical Institute, University of Cologne, Cologne, Germany. <sup>10</sup>Department of Electrical and Computer Engineering, University of California, San Diego, San Diego, CA, USA. <sup>11</sup>Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany. <sup>12</sup>Institute of Agricultural and Nutritional Sciences, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany. <sup>13</sup>BIOSS Centre for Biological Signalling Studies, University of Freiburg, Freiburg, Germany. <sup>14</sup>Plant Cell Biology, Faculty of Biology, University of Marburg, Marburg, Germany. <sup>15</sup>UF Biodiversity Institute, and UF Genetics Institute, University of Florida, Gainesville, FL, USA. <sup>16</sup>New York Botanical Garden, New York, NY, USA. <sup>17</sup>Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plön, Germany. <sup>18</sup>Negaunee Institute for Plant Conservation Science and Action, Chicago Botanic Garden, Glencoe, IL, USA. <sup>19</sup>Program in Plant Biology and Conservation, Northwestern University, Evanston, IL, USA. <sup>20</sup>Department of Horticulture, Michigan State University, East Lansing, MI, USA. <sup>21</sup>Department of Botany, University of Wisconsin-Madison, Madison, WI, USA. <sup>22</sup>Ocean University of China, Qingdao, China. <sup>23</sup>Department of Biological Sciences, University of North Texas, Denton, TX, USA. <sup>24</sup>Centre for Plant Biotechnology and Genomics (CBGP, UPM-INIA), Madrid, Spain. <sup>25</sup>Department of Biodiversity and Conservation, Real Jardín Botánico (RJB-CSIC), Madrid, Spain. <sup>26</sup>Jodrell Laboratory, Royal Botanic Gardens, Kew, London, UK. <sup>27</sup>School of Marine Sciences, University of Maine, Orono, ME, USA. <sup>28</sup>Royal Botanic Garden Edinburgh, Edinburgh, UK. <sup>29</sup>Department of Plant Biology, Laval University, Quebec, Quebec, Canada. <sup>30</sup>Donald Danforth Plant Science Center, St Louis, MO, USA. <sup>31</sup>School of Biological Sciences, The University of Western Australia, Perth, Western Australia, Australia. <sup>32</sup>Kings Park and Botanic Garden, Department of Biodiversity, Conservation and Attractions, Perth, Western Australia, Australia. <sup>33</sup>Australian Tropical Herbarium, James Cook University, Cairns, Queensland, Australia. <sup>34</sup>Department of

Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA. <sup>35</sup>Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK. <sup>36</sup>Umeå Plant Science Centre, Umeå Universitet, Umeå, Sweden. <sup>37</sup>Australian Centre for Evolutionary Biology and Biodiversity, Environment Institute, School of Earth and Environmental Science, University of Adelaide, Adelaide, South Australia, Australia. <sup>38</sup>Department of Plant Sciences, University of Cambridge, Cambridge, UK. <sup>39</sup>Royal Botanic Garden Sydney, Sydney, New South Wales, Australia. <sup>40</sup>Department of Plant Sciences, University of Tennessee, Knoxville, TN, USA. <sup>41</sup>Center for Agricultural Synthetic Biology, University of Tennessee, Knoxville, TN, USA. <sup>42</sup>Department of Food Science, University of Tennessee, Knoxville, TN, USA. <sup>43</sup>Department of Food, Bioprocessing and Nutrition Sciences, North Carolina State University, Raleigh, NC, USA. <sup>44</sup>Institute for Exact and Applied Sciences, University of New Caledonia, Noumea, New Caledonia. <sup>45</sup>Department of Biology, University of Massachusetts, Amherst, MA, USA. <sup>46</sup>USDA-Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, Ames, IA, USA. <sup>47</sup>Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA. <sup>48</sup>Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, UK. <sup>49</sup>Department of Plant Biology, University of Illinois, Urbana-Champaign, Urbana, IL, USA. <sup>50</sup>Fairy Lake Botanical Garden, Chinese Academy of Sciences, Shenzhen, China. <sup>51</sup>Yale-NUS College, Singapore, Republic of Singapore. <sup>52</sup>School of Molecular Sciences, The University of Western Australia, Perth, Western Australia, Australia. <sup>53</sup>Department of Plant Biology, Southern Illinois University, Carbondale, IL, USA. <sup>54</sup>Department of Anatomy, University of Otago, Dunedin, New Zealand. <sup>55</sup>Biology Department, Pennsylvania State University, University Park, PA, USA. <sup>56</sup>Department of Biological Science, California State University Fullerton, Fullerton, CA, USA. <sup>57</sup>Atlanta Botanical Garden, Atlanta, GA, USA. <sup>58</sup>Massey University, School of Fundamental Sciences, Palmerston North, New Zealand. <sup>59</sup>Department of Biology, University of Washington, Seattle, WA, USA. <sup>60</sup>Centre for Plant Sciences, Faculty of Biological Sciences, University of Leeds, Leeds, UK. <sup>61</sup>Department of Plant Sciences, University of Oxford, Oxford, UK. <sup>62</sup>School of Biological Sciences, Monash University, Melbourne, Victoria, Australia. <sup>63</sup>Institut Agronomique néo-Calédonien (IAC), Equipe ARBOREAL, Païta, New Caledonia. <sup>64</sup>Facultad de Ciencias Forestales y de la Conservación de la Naturaleza, Universidad de Chile, Santiago, Chile. <sup>65</sup>Genome Institute of Singapore, Singapore, Singapore. <sup>66</sup>Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany. <sup>67</sup>Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter, UK. <sup>68</sup>Marine Biological Association, The Laboratory, Plymouth, UK. <sup>69</sup>Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA. <sup>70</sup>Department of Biology, University of Toronto Mississauga, Mississauga, Ontario, Canada. <sup>71</sup>School of Life Sciences, La Trobe University, Bundoora, Victoria, Australia. <sup>72</sup>CyVerse, BIO5 Institute, University of Arizona, Tucson, AZ, USA. <sup>73</sup>School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, UK. <sup>74</sup>University of Missouri, St Louis, St Louis, MO, USA. <sup>75</sup>Centre for Organismal Studies Heidelberg, Department of Biodiversity and Plant Systematics, Botanic Garden and Herbarium Heidelberg, University of Heidelberg, Heidelberg, Germany. <sup>76</sup>The Field Museum, Chicago, IL, USA. <sup>77</sup>Royal Horticultural Society Garden Wisley, Woking, UK. <sup>78</sup>University of Reading Herbarium, School of Biological Sciences, University of Reading, Reading, UK. <sup>79</sup>Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden. <sup>80</sup>School of Biological and Chemical Sciences, Queen Mary University of London, London, UK. <sup>81</sup>Boyce Thompson Institute, Cornell University, Ithaca, NY, USA. <sup>82</sup>Environment Institute, School of Biological Science, University of Adelaide, Adelaide, South Australia, Australia. <sup>83</sup>Department of Biology, Duke University, Durham, NC, USA. <sup>84</sup>Smithsonian Environmental Research Center, Edgewater, MD, USA. <sup>85</sup>Department of Biological Sciences, University of Alabama, Tuscaloosa, AL, USA. <sup>86</sup>School of Molecular and Cell Biology, University of the Witwatersrand, Johannesburg, South Africa. <sup>87</sup>Department of Ecology, Evolution and Organismal Biology, Kennesaw State University, Kennesaw, GA, USA. <sup>88</sup>Flower Diversity Institute, Arvada, CO, USA. <sup>89</sup>CSIRO Agriculture and Food, Perth, Western Australia, Australia. <sup>90</sup>Millennium Seed Bank, Wakehurst, Royal Botanic Gardens, Kew, Ardingly, UK. <sup>91</sup>The UWA Institute of Agriculture, The University of Western Australia, Perth, Western Australia, Australia. <sup>92</sup>Centers for Disease Control and Prevention, Atlanta, GA, USA. <sup>93</sup>Department of Primary Industries and Regional Development, Perth, Western Australia, Australia. <sup>94</sup>Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA. <sup>95</sup>Department of Ecology and Evolutionary Biology, The University of Toronto, Ontario, Canada. <sup>96</sup>Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA, USA. <sup>97</sup>Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, USA. <sup>98</sup>Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, USA. <sup>99</sup>Manoa, Honolulu, HI, USA. <sup>100</sup>Department of Biology, University of Louisiana at Lafayette, Lafayette, LA, USA. <sup>101</sup>Morris Arboretum of the University of Pennsylvania, Philadelphia, PA, USA. <sup>102</sup>Koffler Scientific Reserve, University of Toronto, King City, Ontario, Canada. <sup>103</sup>Department of Systematic and Evolutionary Botany, University of Zurich, Zurich, Switzerland. <sup>104</sup>National Taiwan University, Institute of Ecology and Evolutionary Biology, Department of Life Science, Taipei, Taiwan. <sup>105</sup>Systematic Biology, Department of Organismal Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden. <sup>106</sup>Department of Plant Pathology, Physiology and Weed Science, Virginia Tech, Blacksburg, VA, USA. <sup>107</sup>CAS Key Laboratory of Genome Sciences and Information, Beijing Key Laboratory of Genome and Precision Medicine Technologies, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China. <sup>108</sup>Key Laboratory of Agricultural Biological Functional Genes, Northeast Agricultural University, Harbin, China. <sup>109</sup>College of Life Science, Qingdao Agricultural University, Qingdao, China. <sup>110</sup>Agriculture and Agri-Food Canada, Lacombe, Alberta, Canada. <sup>111</sup>Department of Environment and Agriculture, Curtin University, Bentley, Western Australia, Australia. <sup>112</sup>Bond Life Sciences Center, Division of Biological Sciences, University of Missouri, Columbia, MO, USA. <sup>113</sup>Department of Zoology, University of British Columbia, Vancouver, British

Columbia, Canada. <sup>114</sup>University Herbarium and Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, USA. <sup>115</sup>Beijing Genomics Institute-Wuhan, Wuhan, China. <sup>116</sup>BGI-Shenzhen, Shenzhen, China. <sup>117</sup>Agricultural Genome Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. <sup>118</sup>Huahan Gene, Shenzhen, China. <sup>119</sup>MGI, BGI-Shenzhen, Shenzhen, China. <sup>120</sup>Allwegene Technology, Beijing, China. <sup>121</sup>iCarbonX, Shenzhen, China. <sup>122</sup>Georgia Advanced Computing Resource Center, University of Georgia, Athens, GA, USA. <sup>123</sup>Department of Biological Sciences, Western Michigan University, Kalamazoo, MI, USA. <sup>124</sup>Department of Computer Science and Engineering, University of California, San Diego, San Diego, CA, USA. <sup>125</sup>Lawrence J. Ellison Institute for Transformative Medicine, University of Southern California, Los Angeles, CA, USA. <sup>126</sup>Microbiology, Immunology and Biochemistry, The University of Tennessee Health Science Center, Memphis, TN, USA. <sup>127</sup>Department of Computer Science, University of Illinois, Urbana-Champaign, Urbana, IL, USA. <sup>128</sup>Division of Biological Sciences, University of Missouri, Columbia, MO, USA. <sup>129</sup>Arizona Research Laboratories, University of Arizona, Tucson, AZ, USA. <sup>130</sup>Key Laboratory of Plant

Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, China. <sup>131</sup>Flower Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, China. <sup>132</sup>Department of Genetics, Matthias Schleiden Institute, Friedrich-Schiller-University Jena, Jena, Germany. <sup>133</sup>Department of Medicine, University of Alberta, Edmonton, Alberta, Canada. <sup>134</sup>These authors contributed equally: Eric J. Carpenter, Matthew A. Gitzendanner, Zheng Li, Siavash Mirarab, Martin Porsch, Kristian K. Ullrich, Lisa DeGironimo, Patrick P. Edger, Ingrid E. Jordon-Thaden, Steve Joya, Tao Liu, Barbara Melkonian, Nicholas W. Miles, Lisa Pokorny Montero, Charlotte Quigley, Philip Thomas, Juan Carlos Villarreal. <sup>135</sup>These authors jointly supervised this work: James H. Leebens-Mack, Michael S. Barker, Michael K. Deyholos, Sean W. Graham, Ivo Grosse, Michael Melkonian, Siavash Mirarab, Marcel Quint, Stefan A. Rensing, Douglas E. Soltis, Pamela S. Soltis, Dennis W. Stevenson, Claude W. dePamphilis, Mark W. Chase, J. Chris Pires, Carl J. Rothfels, Jun Yu, Yong Zhang, Tandy Warnow, Shuangxiu Wu, Gane Ka-Shu Wong. \*e-mail: jleebensmack@uga.edu; gane@ualberta.ca

# Article

## Methods

### Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized, although simulations included in the genome duplication analyses did include drawing from random distributions. The investigators were not blinded to allocation during experiments and outcome assessment.

### Transcriptome sequencing

RNA was isolated from young vegetative tissue from all of the species that were included in our phylogenomic analyses as described elsewhere<sup>39,67,68</sup>. Reproductive tissues were also included for some species (Supplementary Table 1). Transcript assembly, contaminant identification and gene-family circumscription were also performed as described previously<sup>39</sup> and are described in more detail in the Supplementary Methods.

### Phylogeny reconstruction

Analyses were performed on single-copy gene trees using ASTRAL to account for variation among gene trees owing to incomplete lineage sorting<sup>15,69</sup>. ASTRAL analyses were performed on gene trees estimated from unbinned amino acid alignments, first and second codons, statistically binned supergenes with unweighted bins<sup>70,71</sup> and filtered taxon sets (excluding 'rogue' taxa as described below), with filtering of gene-tree bootstrap support thresholds of up to 33% to see whether the effects of gene-tree estimation error could be reduced (Supplementary Fig. 6). Binning left the majority of genes in singleton bins and had minimal effects on the overall species tree. Unless otherwise specified, we use 'ASTRAL topology' to refer to the tree inferred from 410 unbinned amino acid alignments in which branches with 33% or less support are contracted. In addition, supermatrix analyses were performed on concatenated nuclear gene alignments and concatenated plastid gene alignments compiled using previously described methods<sup>72</sup>. All scripts used to perform analyses on the nuclear gene data are available at <https://doi.org/10.5281/zenodo.3255100>.

**Multiple sequence alignment and data filtering.** We built a multiple sequence alignment based on predicted amino acid sequences of each gene and forced DNA sequences to conform to the amino acid alignment. We first divided sequences in each gene into two subsets, full-length and abnormal sequences, and then used PASTA<sup>73</sup> with default settings to align full-length sequences and UPP<sup>74</sup> to add abnormal sequences to the full-length alignment. We designated as abnormal any sequence that was 66% shorter or 66% longer than the median length of the full-length gene sequences. Once UPP alignments were obtained, we removed from them all unaligned (that is, insertion) sites. DNA alignments were then derived from amino acid sequence alignments (FAA2FNA) and third codon positions were removed owing to extreme among-species variation in GC content (Supplementary Fig. 7). To reduce running time, we then masked all sites from the alignment that contained more than 90% gaps. Finally, because the inclusion of fragmentary data in gene-tree estimation can be problematic<sup>75</sup>, we removed any sequence that had a gap for at least 67% of the sites in the site-filtered alignment (the 67% threshold was chosen based on simulation results<sup>75</sup>). Gene sequence occupancy for 410 single-copy genes in the 1,178 accessions used in our analyses is displayed as a frequency histogram (Supplementary Fig. 4) and a heat map (Supplementary Fig. 5).

In addition to filtering gappy sites and fragmentary sequences, we identified and removed sequences that were placed on extremely long branches on their respective gene trees. To identify these, we used the initial alignments to build gene trees (see below). We then rooted each gene tree by finding the bipartition that separated the largest exclusive group of outgroup or red algae taxa. If red algae were entirely missing for the gene, we used Glaucophyta, Prasinococcales, prasinophytes,

*Volvox carteri*, *Chlamydomonas reinhardtii* or *Klebsormidium nitens*. We then removed any sequences that had a root-to-tip distance that was four standard deviations longer than the median root-to-tip distance in each gene tree. Once these sequences on long branches were removed, alignments were re-estimated using the same approach described above, and new gene trees were estimated.

**Gene-tree estimation.** To estimate gene trees, we used RAXML v.8.1.17<sup>76</sup>, with one starting tree for building initial trees (used for long-branch filtering) and 10 different starting trees for final gene trees. Support was assessed with 100 replicates of bootstrapping. For DNA analyses, the GTR substitution model and the GAMMA-distributed site rates were used. For amino acid sequences, we used a Perl script adapted from the RAXML website to search among 16 different substitution models on a fixed starting tree per gene and chose the model with the highest likelihood (JTT, JTTF or JTTDCMUT were selected for 349 out of 410 genes). For amino acid trees, we also used the GAMMA-distributed site rates.

**Species tree estimation.** We used ASTRAL-II<sup>15</sup> v.5.0.3 to estimate the species tree on the basis of all 410 genes; using 384 genes that each included at least half of the species changed only 3 low-support branches. We used multi-locus bootstrapping<sup>77,78</sup> and the built-in local posterior probabilities of ASTRAL to estimate branch support<sup>69</sup> and to test for polytomies<sup>79</sup>, drawn on species trees estimated based on the maximum-likelihood gene trees. We also used the built-in functionality of ASTRAL (version 4.11.2) to compute the percentage of gene trees that agreed with each branch in the species tree, by finding the average number of gene-tree quartets defined around the branch (choosing one taxon from each side) that were congruent with the species tree and used DiscoVista<sup>80</sup> to visualize them (Fig. 4). Median representation of each species across the 410 single-copy gene trees was 82.4% with 88.2% and 67.1% of species having assemblies for at least 50% or 75% of the 410 single-copy genes, respectively. A large body of work on phylogenetic methodologies has established that gene and species tree estimation can be robust to missing data, particularly with dense taxon sampling<sup>75,81,82</sup>. Recent papers have even established statistical consistency under missing data<sup>83</sup>. Similar evidence of robustness also exists in the context of concatenated analyses<sup>84–86</sup>.

All supermatrix analyses are based on the filtered amino acid and first and second codon position alignments that included at least half of the species for 384 genes. The (1) unfiltered supermatrices used the gene alignments as is; the (2) eudicot supermatrices retained only eudicot species in the supermatrix; and the (3) supermatrices with eight 'rogue' taxa removed (*Dillenia indica*, *Tetrastigma obtectum*, *Tetrastigma voinierianum*, *Vitis vinifera*, *Cissus quadrangularis*, 'Spirotaenia' sp., *Ceratophyllum demersum* and *Prasinococcus capsulatus*) that varied in placement among our full ASTRAL, supermatrix and plastid genome analyses. Well-supported branching orders were stable among analyses (Supplementary Fig. 6).

Maximum-likelihood supermatrix analyses were performed using ExaML v.3.0.14<sup>87</sup>. Similar to the gene-tree analyses, the GAMMA model of rate heterogeneity across sites was used for all maximum-likelihood supermatrix analyses. To better handle model heterogeneity across genes, we divided the supermatrix into partitions. For the amino acid alignments, the protein model selected for each gene family in the gene-tree estimation process was used to group genes into partitions, creating one partition per substitution model. For the nucleotide alignments, we estimated the GTR transition rate parameters and the alpha shape parameter for each codon position (first and second positions) of each alignment using RAXML v.8.1.21<sup>76</sup>. We then projected the maximum-likelihood parameter values for each gene into a two-dimensional plane using principal component analysis<sup>88</sup>. We performed *k*-means clustering<sup>89</sup> in  $R^{90}$  to group the codon positions into partitions, selecting *k* = 8, which accounted for 80% of the variation. Trees derived from nucleotide alignments can be found at <https://doi.org/10.5281/zenodo.3255100>.

To examine the influence of the starting tree on the likelihood of the final tree, we performed preliminary analyses on an earlier version of our supermatrices. We generated nine different maximum-parsimony starting trees using RAxML v.8.1.21 and one maximum-likelihood starting tree using FastTree-2 v.2.1.5<sup>91</sup>. We then ran ExaML on each of the starting trees, noting the final maximum-likelihood score. We found that in all cases, the ExaML maximum-likelihood tree using the FastTree-2 maximum-likelihood starting tree had a better maximum-likelihood score than any of the ExaML maximum-likelihood trees using maximum-parsimony starting trees. Thus, for all of the supermatrix analyses, we used FastTree-2 to generate our initial starting tree. Support was inferred for the branches of the final tree from 100 bootstrap replicates.

Outgroup taxa from outside Archaeplastida were used to root all species trees estimated using nuclear genes (all ASTRAL and supermatrix analyses). The plastome supermatrix tree for Viridiplantae was rooted using Rhodophyta as outgroup.

### Inferring and placing WGDs

#### DupPipe analyses of WGDs from transcriptomes of single species.

For each transcriptome, we used the DupPipe pipeline to construct gene families and estimate the age distribution of gene duplications<sup>16,17</sup>. We translated DNA sequences and identified reading frames by comparing the Genewise<sup>92</sup> alignment to the best-hit protein from a collection of proteins from 25 plant genomes from Phytozome<sup>93</sup>. For all DupPipe runs, we used protein-guided DNA alignments to align our nucleic acid sequences while maintaining the reading frame. We estimated synonymous divergence ( $K_s$ ) using PAML with the F3X4 model<sup>94</sup> for each node in the gene-family phylogenies. We identified peaks of gene duplication as evidence of ancient WGDs in histograms of the age distribution of gene duplications ( $K_s$  plots). We identified species with potential WGDs by comparing their paralogue age distribution to a simulated null using a Kolmogorov–Smirnov goodness of fit test<sup>95</sup>. We then used mixture modelling and manual curation to identify significant peaks consistent with a potential WGD and to estimate their median paralogue  $K_s$  values. Significant peaks were identified using a likelihood ratio test in the boot.comp function of the package mixtools in R<sup>96</sup>.

**Estimating orthologous divergence.** To place putative WGDs in relation to lineage divergence, we estimated the synonymous divergence of orthologues among pairs of species that may share a WGD based on their phylogenetic position and evidence from the within-species  $K_s$  plots. We used the RBH Orthologue pipeline<sup>17</sup> to estimate the mean and median synonymous divergence of orthologues and compared those to the synonymous divergence of inferred paleopolyploid peaks. We identified orthologues as reciprocal best blast hits in pairs of transcriptomes. Using protein-guided DNA alignments, we estimated the pairwise synonymous divergence for each pair of orthologues using PAML with the F3X4 model<sup>94</sup>. WGDs were interpreted to have occurred after lineage divergence if the median synonymous divergence of WGD paralogues was younger than the median synonymous divergence of orthologues. Similarly, if the synonymous divergence of WGD paralogues was older than that orthologue synonymous divergence, then we interpreted those WGDs as shared.

#### MAPS analyses of WGDs from transcriptomes of multiple species.

To infer and locate putative WGDs in our datasets, we used a gene-tree sorting and counting algorithm, the multi-taxon paleopolyploidy search (MAPS) tool<sup>18</sup>. For each MAPS analysis, we selected at least two species that potentially share a WGD in their ancestry as well as representative species from lineages that may phylogenetically bracket the WGD. MAPS uses this given species tree to filter collections of nuclear gene trees for subtrees consistent with relationships at each node in the species tree. Using this filtered set of subtrees, MAPS identifies and records nodes with a gene duplication shared by descendant taxa. To infer and locate a potential WGD, we compared the number of duplications observed

at each node to a null simulation of background gene birth and death rates<sup>97,98</sup>. A Fisher's exact test, implemented in R<sup>90</sup>, was used to identify locations with significant increases in gene duplication compared with a null simulation (Supplementary Table 5). Locations with significantly more duplications than expected were then compared to a simulated WGD at this location. If the observed duplications were consistent with this simulated WGD using Fisher's exact test, we identified the location as a WGD if it was consistent with inferences from  $K_s$  plots and orthologue divergence data. In some cases, MAPS inferred significant duplications without apparent signatures in  $K_s$  plots or previously published research. In these cases, we recognized the event as a significant burst of gene duplication.

Each MAPS analysis was designed to place focal WGDs near the centre of a species tree to minimize errors in WGD inference. Errors in transcriptome or genome assembly, gene-family clustering and the construction of gene-family phylogenies can result in topological errors in gene trees<sup>99</sup>. Previous studies have suggested that errors in gene trees can lead to biased placements of duplicates towards the root of the tree and losses towards the tips of the tree<sup>100</sup>. For this reason, we aimed to put focal nodes for a particular MAPS analysis test in the middle of the phylogeny. To further decrease potential error in our inferences of gene duplications, we required at least 45% of the ingroup taxa to be present in all subtrees analysed by MAPS<sup>97</sup>. If this minimum requirement of ingroup taxa numbers is not met, the gene subtree will be filtered out and excluded from our analysis. Increasing taxon occupancy leads to a more accurate inference of duplications and reduces some of the biases in mapping duplications onto a species tree<sup>100,101</sup>. To maintain sufficient gene-tree numbers for each MAPS analysis, we used collections of gene-family phylogenies for six to eight taxa to infer ancient WGDs.

For each MAPS analysis, the transcriptomes were translated into amino acid sequences using the TransPipe pipeline<sup>17</sup>. Using these translations, we performed reciprocal protein BLAST (BLASTp) searches among datasets for the MAPS analysis using a cut-off of  $E = 1 \times 10^{-5}$ . We clustered gene families from these BLAST results using OrthoFinder under the default parameters<sup>102</sup>. Using a custom Perl script (<https://bitbucket.org/barkerlab/MAPS>), we filtered for gene families that contained at least one gene copy from each taxon in a given MAPS analysis and discarded the remaining OrthoFinder clusters. We used PASTA<sup>73</sup> for automatic alignment and phylogeny reconstruction of gene families. For each gene-family phylogeny, we ran PASTA until we reached three iterations without an improvement in likelihood score using a centroid breaking strategy. Within each iteration of PASTA, we constructed subset alignments using MAFFT<sup>103</sup>, used Muscle<sup>104</sup> for merging these subset alignments and RAxML<sup>76</sup> for tree estimation. The parameters for each software package were the default options for PASTA (<https://bitbucket.org/barkerlab/1kp>). We used the best-scoring PASTA tree for each multi-species nuclear gene family to collectively estimate the numbers of shared gene duplications on each branch of the given species.

To generate null simulations, we first estimated the mean background gene duplication rate ( $\lambda$ ) and gene loss rate ( $\mu$ ) with WGDgc<sup>98</sup> (Supplementary Tables 5, 11). Gene count data were obtained from OrthoFinder<sup>102</sup> clusters associated with each species tree (Supplementary Table 5).  $\lambda$  and  $\mu$  were estimated using only gene clusters that spanned the root of their respective species trees, which has been shown to reduce biases in the maximum-likelihood estimates<sup>98</sup> of  $\lambda$  and  $\mu$ . We chose a maximum gene-family size of 100 for parameter estimation, which was necessary to provide an upper bound for numerical integration of node states<sup>98</sup>. We provided a prior probability distribution on the number of genes at the root of each species tree, such that ancestral gene-family sizes followed a shifted geometric distribution with mean equal to the average number of genes per gene family across species (Supplementary Table 5).

Gene trees were then simulated within each MAPS species trees using the GuestTreeGen program from GenPhyloData<sup>105</sup>. For each species tree, we simulated 3,000 gene trees with at least one tip per species; 1,000 gene trees at the  $\lambda$  and  $\mu$  maximum-likelihood estimates, 1,000

# Article

gene trees at half the estimated  $\lambda$  and  $\mu$ , and 1,000 trees at three times  $\lambda$  and  $\mu$ . For all simulations, we applied the same empirical prior used for estimation of  $\lambda$  and  $\mu$ . We then randomly resampled 1,000 trees without replacement from the total pool of gene trees 100 times to provide a measure of uncertainty on the percentage of subtrees at each node. For positive simulations of WGDs, we simulated gene trees using the same approach used to generate null distributions (Supplementary Table 5) but incorporated a WGD at the test branch. Previous empirical estimates of paralogues retained following a plant WGD are 10% on average<sup>106</sup>. To be conservative for inferring WGDs in our MAPS analyses, we allowed at least 20% of the genes to be retained following the simulated WGD to account for biased gene retention and loss. For WGDs that might have a lower gene retention rate, we used an additional simulation using 15% gene retention (Supplementary Table 6).

## Gene-family evolution

**Transcriptome-based gene-family size estimation.** To robustly estimate gene-family sizes from transcriptomic data, we needed to overcome three major challenges: (1) the fragmentation of transcript sequences; (2) the absence of low-abundance transcripts; and (3) the over-prediction of gene-family sizes due to assembly duplications and biological isoforms. We dealt with these challenges as follows.

**Fragmentation of data.** The multiple sequence alignments used to construct the domain-specific profile hidden Markov models (HMMs) ranged from 23 to 463 amino acids in length; 78% of these alignments were shorter than 120 amino acids, and 84.6% of the assembled and translated transcripts were longer than 120 amino acids. By mainly characterizing gene families using single domains (Supplementary Table 9), we limited the effect of the fragmentation of transcripts from the assembly of short read data. HMMs used for gene-family classification and decision rules obtained from either published work<sup>107</sup> or gene-family experts are given in Supplementary Table 9; 12 out of 23 gene families were classified by a single 'should' rule, 2 out of 23 were defined by a XOR 'should' rule, which also classifies a sequence by the presence of a single domain, 8 out of 23 gene families were classified by a more complex rule set including 'should not' rules. The only gene family for which multiple domains needed to be present was the PLS subfamily of the PPR gene family.

**Loss of low abundance transcripts.** To account for possible bias in the sampling of the gene space, all species that showed low levels of transcriptome completeness were removed. The lowest value of transcriptome completeness obtained from 30 annotated plant genomes was used as the lower exclusion limit. We removed all samples in which more than 42.5% of BUSCO<sup>31</sup> sequences were missing using default settings and the eukaryotic dataset as the query database.

**Gene-family over-prediction.** We clustered assembled protein sequences by sequence similarity and merged sequences that showed at least 99% identity. To check for the possibility of merging sequences that should be counted separately, different identity cut-offs were compared between the IKP datasets and 32 annotated plant genomes.

Extended Data Figure 3c, d shows the average gene-family sizes for 23 gene families and 13 clades obtained from IKP samples and 32 annotated plant genomes. These gene-family sizes show a high Pearson correlation ( $r = 0.95$ ) between IKP samples and plant genomes, and therefore a linear relationship between the two approaches is indicated. The results from the IKP dataset are on average smaller by a factor of 2.3. Although this is a clear underestimate, the scale factor by which the estimate is too small is relatively consistent, especially as the gene-family sizes increase.

**Sequence clustering.** We used cdhit v.4.5.7<sup>108,109</sup> to reduce the number of protein sequence duplications in the dataset. We assessed 100%, 99.5%, 99%, 95% and 90% sequence identity thresholds. The percentage of remaining sequences for the IKP samples and 32 reference genomes

is displayed in Extended Data Fig. 3f. We chose 99% sequence identity as the value to use for this study.

**Estimation of gene-family size.** Gene-family experts provided the knowledge to classify protein sequences as members of gene families with profile HMMs. In total, 46 HMMs representing 23 large gene families<sup>30</sup> were used to estimate gene-family sizes in the analysed species. Classification rules and HMMs for 14 gene families that have been published previously<sup>107</sup> were converted to HMMER3 format and used in this study. Gene-family classification rules and HMMs for the remaining nine families can be found in Supplementary Table 8. HMMs were taken from the Pfam database (accessed 12 May 2016) or were provided by gene-family experts (Supplementary Table 8). HMMER<sup>110</sup> (v.3.1b2) was used to scan for matches in the filtered IKP dataset. Where available, gathering thresholds were used; otherwise an *E*-value cut-off was applied to indicate domain presence. If the *E* value is not noted in Supplementary Table 9, the default *E* value of 10 was applied. The results on the species level are listed in Supplementary Table 10s.

**Statistical test for expansions and contractions.** To assess whether a gene family expanded or contracted in a lineage, we compared a weighted average of gene numbers in adjacent clades and grades (Fig. 4). We also checked for expansions and contractions within clades but did not find any statistically significant shifts. The counts of gene-family members from two clades or grades were compared with a Kolmogorov–Smirnov test with a *P*-value threshold of  $1 \times 10^{-6}$  in R<sup>90</sup>. The tests conducted in this study are listed in Supplementary Table 7. Fold changes were computed using the trimmed arithmetic mean in which the top and bottom 5% of the data were discarded. Only expansions larger than 1.5 fold (or contractions smaller than 2/3) are reported.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

All raw sequence reads have been posted in the NCBI SRA database under BioProject accession PRJEB4922. SRA entries for each assembly are listed in Supplementary Table 1. All sequence, gene tree and species tree data can be accessed through CyVerse Data Commons at <https://doi.org/10.25739/8m7t-4e85>. In addition, gene-family nucleotide and amino acid FASTA files can also be found at <http://jlmwiki.plantbio.uga.edu/onekp/v2/>; multiple sequence alignments, gene trees and species trees for single-copy nuclear genes included in phylogenomic analyses are also at <https://doi.org/10.5281/zenodo.3255100>;  $K_s$  plots, alignments and trees used for WGD analyses can be found at <https://bitbucket.org/barkerlab/1kp>; and data used for gene-family expansion analyses can be found at <https://github.com/GrosseLab/OneKP-gene-family-evo>.

## Code availability

Scripts used for phylogenomic species tree analyses are available at <https://doi.org/10.5281/zenodo.3255100>. Scripts used for MAPS analyses of WGDs are available at <https://bitbucket.org/barkerlab/maps> and scripts used for gene-family expansion analyses are available at <https://github.com/GrosseLab/OneKP-gene-family-evo>. All script files are also accessible through CyVerse Data Commons at <https://doi.org/10.25739/8m7t-4e85>.

67. Johnson, M. T. J. et al. Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. *PLoS ONE* **7**, e50226 (2012).

68. Jordon-Thaden, I. E., Chanderbali, A. S., Gitzendanner, M. A. & Soltis, D. E. Modified CTAB and TRIzol protocols improve RNA extraction from chemically complex Embryophyta. *Appl. Plant Sci.* **3**, 1400105 (2015).



69. Sayyari, E. & Mirarab, S. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654–1668 (2016).
70. Mirarab, S., Bayzid, M. S., Boussau, B. & Warnow, T. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* **346**, 1250463 (2014).
71. Bayzid, M. S., Mirarab, S., Boussau, B. & Warnow, T. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PLoS ONE* **10**, e0129183 (2015).
72. Gitzendanner, M. A., Soltis, P. S., Wong, G. K.-S., Ruhfel, B. R. & Soltis, D. E. Plastid phylogenomic analysis of green plants: a billion years of evolutionary history. *Am. J. Bot.* **105**, 291–301 (2018).
73. Mirarab, S., Nguyen, N. & Warnow, T. PASTA: ultra-large multiple sequence alignment. *Res. Comput. Mol. Biol.* **8394**, 177–191 (2014).
74. Nguyen, N.-P. D., Mirarab, S., Kumar, K. & Warnow, T. Ultra-large alignments using phylogeny-aware profiles. *Genome Biol.* **16**, 124 (2015).
75. Sayyari, E., Whitfield, J. B. & Mirarab, S. Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. *Mol. Biol. Evol.* **34**, 3279–3291 (2017).
76. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
77. Seo, T.-K., Kishino, H. & Thorne, J. L. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. *Proc. Natl Acad. Sci. USA* **102**, 4436–4441 (2005).
78. Mirarab, S., Bayzid, M. S. & Warnow, T. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst. Biol.* **65**, 366–380 (2016).
79. Sayyari, E. & Mirarab, S. Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes* **9**, 132 (2018).
80. Sayyari, E., Whitfield, J. B. & Mirarab, S. DiscoVista: interpretable visualizations of gene tree discordance. *Mol. Phylogenet. Evol.* **122**, 110–115 (2018).
81. Dobrin, B. H., Zwickl, D. J. & Sanderson, M. J. The prevalence of terraced trees in analyses of phylogenetic data sets. *BMC Evol. Biol.* **18**, 46 (2018).
82. Molloy, E. K. & Warnow, T. To include or not to include: the impact of gene filtering on species tree estimation methods. *Syst. Biol.* **67**, 285–303 (2018).
83. Nute, M., Chou, J., Molloy, E. K. & Warnow, T. The performance of coalescent-based species tree estimation methods under models of missing data. *BMC Genomics* **19**, 286 (2018).
84. Wiens, J. J. Missing data, incomplete taxa, and phylogenetic accuracy. *Syst. Biol.* **52**, 528–538 (2003).
85. Wiens, J. J. & Morrill, M. C. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst. Biol.* **60**, 719–731 (2011).
86. Lam, V. K. Y. et al. Phylogenomic inference in extremis: a case study with mycoheterotroph plastomes. *Am. J. Bot.* **105**, 480–494 (2018).
87. Kozlov, A. M., Aberer, A. J. & Stamatakis, A. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* **31**, 2577–2579 (2015).
88. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Phil. Mag.* **2**, 559–572 (1901).
89. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: a K-means clustering algorithm. *J. R. Stat. Soc. C* **28**, 100–108 (1979).
90. R Core Team. *R: A Language and Environment for Statistical Computing*. <http://www.R-project.org/> (R Foundation for Statistical Computing, 2014).
91. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
92. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
93. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
94. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
95. Cui, L. et al. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749 (2006).
96. Benaglia, T., Chauveau, D., Hunter, D. & Young, D. mixtools: an R package for analyzing mixture models. *J. Stat. Softw.* **32**, 1–29 (2009).
97. Li, Z. et al. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc. Natl Acad. Sci. USA* **115**, 4713–4718 (2018).
98. Rabier, C.-E., Ta, T. & Ané, C. Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Mol. Biol. Evol.* **31**, 750–762 (2014).
99. Yang, Y. & Smith, S. A. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics* **14**, 328 (2013).
100. Hahn, M. W. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.* **8**, R141 (2007).
101. Smith, S. A., Moore, M. J., Brown, J. W. & Yang, Y. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* **15**, 150 (2015).
102. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
103. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
104. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
105. Sjöstrand, J., Arvestad, L., Lagergren, J. & Sennblad, B. GenPhyloData: realistic simulation of gene family evolution. *BMC Bioinformatics* **14**, 209 (2013).
106. Tiley, G. P., Ané, C. & Burleigh, J. G. Evaluating and characterizing ancient whole-genome duplications in plants with gene count data. *Genome Biol. Evol.* **8**, 1023–1037 (2016).
107. Lang, D. et al. Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol. Evol.* **2**, 488–503 (2010).
108. Li, W., Jaroszowski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).
109. Li, W., Jaroszowski, L. & Godzik, A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* **18**, 77–82 (2002).
110. Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
111. Henschel, K. et al. Two ancient classes of MIKC-type MADS-box genes are present in the moss *Physcomitrella patens*. *Mol. Biol. Evol.* **19**, 801–814 (2002).

**Acknowledgements** The IKP initiative was funded by the Alberta Ministry of Advanced Education and Alberta Innovates AITF/iCORE Strategic Chair (RES0010334) to G.K.-S.W., Musea Ventures, The National Key Research and Development Program of China (2016YFE0122000), The Ministry of Science and Technology of the People's Republic of China (2015BAD04B01/2015BAD04B03), the State Key Laboratory of Agricultural Genomics (2011DQ782025) and the Guangdong Provincial Key Laboratory of core collection of crop genetic resources research and application (2011A091000047). Sequencing activities at BGI were also supported by the Shenzhen Municipal Government of China (CX22014042112021913/JCYJ20150529150409546/JCYJ20150529150505656). Computation support was provided by the China National GeneBank (CNCB), the Texas Advanced Computing Center (TACC), WestGrid and Compute Canada; considerable support, including personnel, computational resources and data hosting, was also provided by the iPlant Collaborative (CyVerse) funded by the National Science Foundation (DBI-1265383), National Science Foundation grants IOS 0922742 (to C.W.d., P.S.S., D.E.S. and J.H.L.-M.), IOS-1339156 (to M.S.B.), DEB 0830009 (to J.H.L.-M., C.W.d., S.W.G. and D.W.S.), EF-0629817 (to S.W.G. and D.W.S.), EF-1550838 (to M.S.B.), DEB 0733029 (to T.W. and J.H.L.-M.), and DBI 1062335 and 1461364 (to T.W.), a National Institutes of Health Grant 1R01DA025197 (to T.M.K., C.W.d. and J.H.L.-M.), Deutsche Forschungsgemeinschaft grants Qu 141/5-1, Qu 141/6-1, GR 3526/7-1, GR 3526/8-1 (to M.Q. and I.G.) and a Natural Sciences and Engineering Research Council of Canada Discovery grant (to S.W.G.). We thank all national, state, provincial and regional resource management authorities, including those of province Nord and province Sud of New Caledonia, for permitting collections of material for this research.

**Author contributions** Framing of research and writing was carried out by J.H.L.-M., M.S.B., E.J.C., M.K.D., M.A.G., S.W.G., I.G., Z.L., M. Melkonian, S.M., M.P., M.Q., S.A.R., D.E.S., P.S.S., D.W.S., K.K.U., N.J.W. and G.K.-S.W. Samples were collected and RNA was prepared by L.D., P.P.E., I.E.J.-T., S.J., T.L., B.M., N.W.M., L.P., C.Q., P.T., J.C.V., M.M.A., M.S.B., M.D.B., R.S.B., D.J.B., R.M.B., E.B., S.F.B., D.O.B., J.N.B., K.P.B., V.B.-S., A.L.C., S.B.C., Z.C., Y.C., C. Chater, J.M.C., T.C., N.D.C., H. Clayton, S. Covshoff, B.J.C.-S., H. Cross, C.W.d., J.P.D., R.D., R.C.D., V.S.D.S., S.E., E.F., N.F., K.J.F., D.A.F., P.M.F., S.K.F., B.F., N.G., G.G., M.A.G., G.T.G., F.Q.Y.G., S. Greiner, A.H., J. M. Heaney, K.E.H., K.H., J. M. Hibberd, R.G.J.H., P.M.H., M.T.J.J., R.J., B.J., M.V.K., K.E., E.A.K., M.A.K., M.V.K., K.K., T.M.K., V.L., A.L., A.R.L., R. Lentz, F.-W.L., A.J.L., M.L., P.S.M., E.M., M.K.M., M. McKain, T.M., J.R.M., R.E.M., M.N.N., Y.P., P.R., D.R., C.W.R., M.R., R.F.S., A.K.S., M.S., E.E.S., E.-M.S., H.S., S.S., E.B.S., A.J.S., S.W.S., E.M.S., C.S., A.G.S., A.S., C.N.S., J.R.S., P.S., J.A.T., H.T., D.T., M.V., C.-N.W., S.G.W., M.W., S. Weststrand, J.H.W., D.F.W., N.J.W., S. Wu, A.S.W., Y.Y., D.Z., C.Z., J.Z., M.W.C., M.K.D., S.W.G., J.H.L.-M., M. Melkonian, J.C.P., C.J.R., D.E.S., P.S.S., D.W.S. and J.Y. (jointly led by M.W.C., M.K.D., S.W.G., J.H.L.-M., M. Melkonian, J.C.P., C.J.R., D.E.S., P.S.S., D.W.S. and J.Y.; major contributions by L.D., P.P.E., I.E.J.-T., S.J., T.L., B.M., N.W.M., L.P., C.Q., P.T. and J.C.V.). RNA sequencing and transcriptome assembly were carried out by E.J.C., C. Chen, L.C., S. Cheng, J.L., R. Li, X.L., H.L., Y.O., X.S., X.T., J.T., Z.T., F.W., J.W., X.W., G.K.-S.W., X.X., Z.Y., F.Y., X.Z., F.Z., Y. Zhu and Y. Zhang. (led by Y. Zhang; major contributions by E.J.C.). Samples were validated and contaminants were filtered by J.Y., S.A., M.S.B., T.J.B., E.J.C., S.W.G., J.H.L.-M., T.L., S.M., N.-p.N., X.S., K.K.U. and S. Wu. (led by S. Wu; major contributions by J.Y.). Gene-family circumscription and phylogenetic analyses were carried out by S.M., N.-p.N., M.A.G., S.A., J.P.D., N.M., D.R.N., E.S., D.E.S., P.S.S., D.W.S., E.K.W., R.L.W., N.J.W., C.W.d., S.W.G., J.H.L.-M. and T.W. (jointly led by S.M., C.W.d., S.W.G., J.H.L.-M. and T.W.; major contributions by S.M.). Genome duplication analyses were carried out by Z.L., H.A., N.A., A.E.B., S. Galuska, S.A.J., T.I.K., H.K., P.-I., H.E.M., X.Q., C.R.R., E.B.S., B.L.S., G.P.T., S.R.W., R.Y., S.Z. and M.S.B. (led by M.S.B.; major contributions by Z.L.). Gene-family expansion analyses were carried out by M.P., K.K.U., L.G., M. Melkonian, D.R.N., G.T., G.K.-S.W., I.G., S.A.R. and M.Q. (jointly led by I.G., S.A.R. and M.Q.; major contributions by M.P. and K.K.U.).

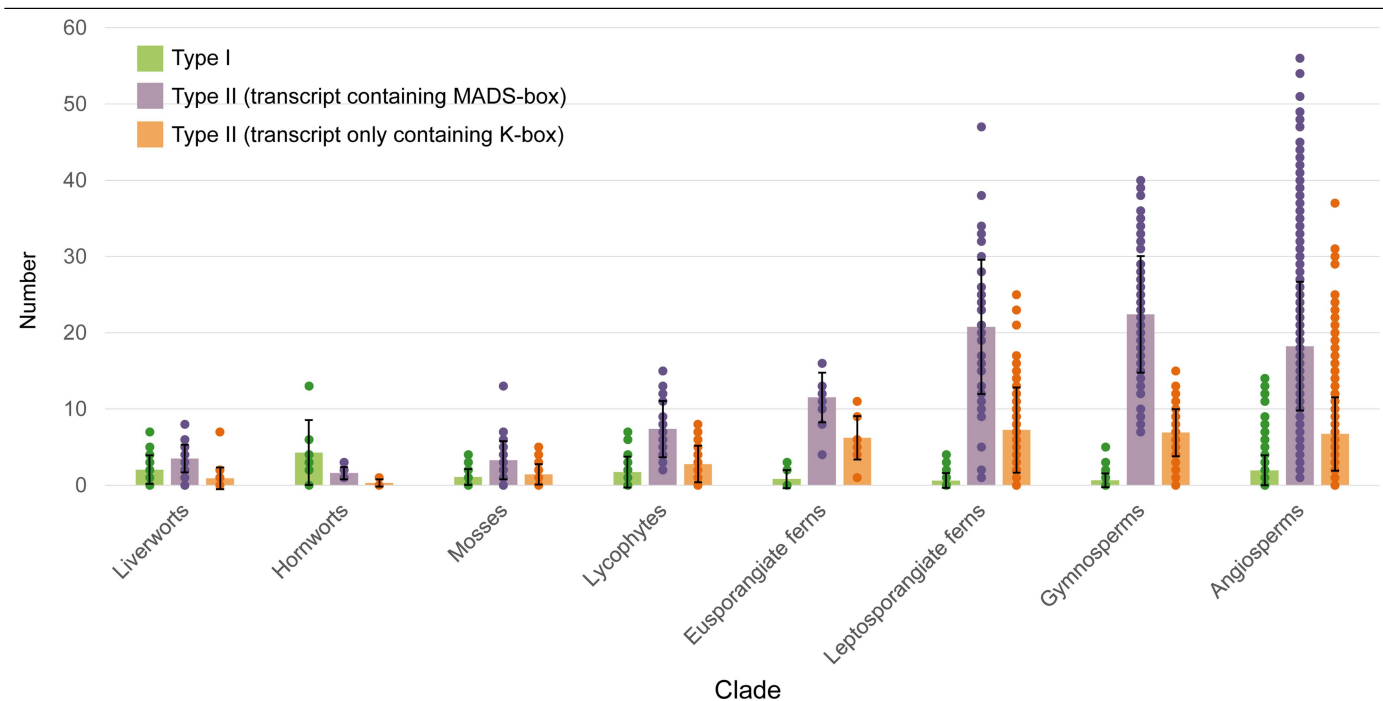
**Competing interests** The authors declare no competing interests.

#### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1693-2>.

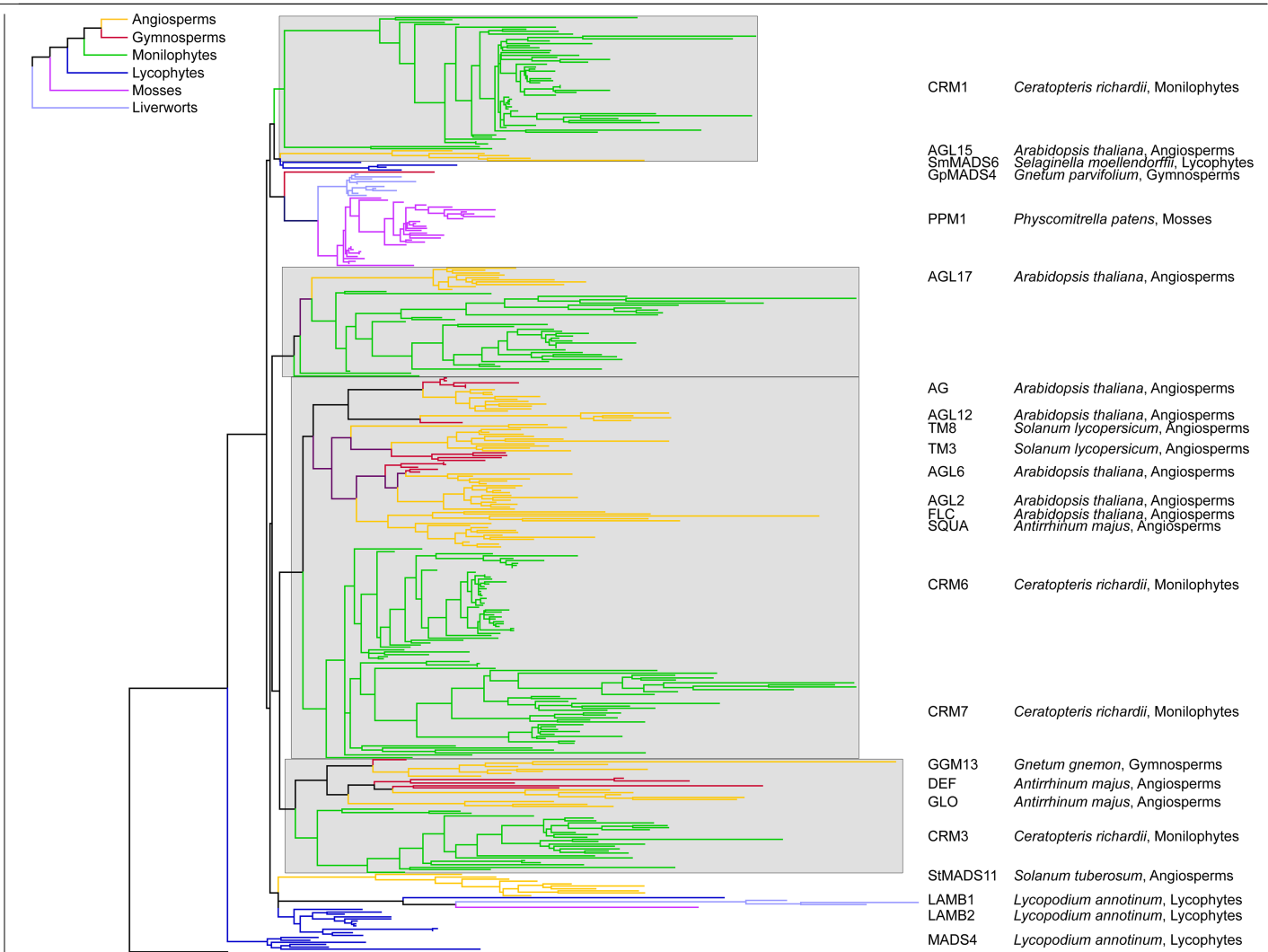
**Correspondence and requests for materials** should be addressed to J.H.L.-M. or G.K.-S.W.  
**Reviewer information** Nature thanks Paul Kenrick, Magnus Nordborg, Patrick Wincker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



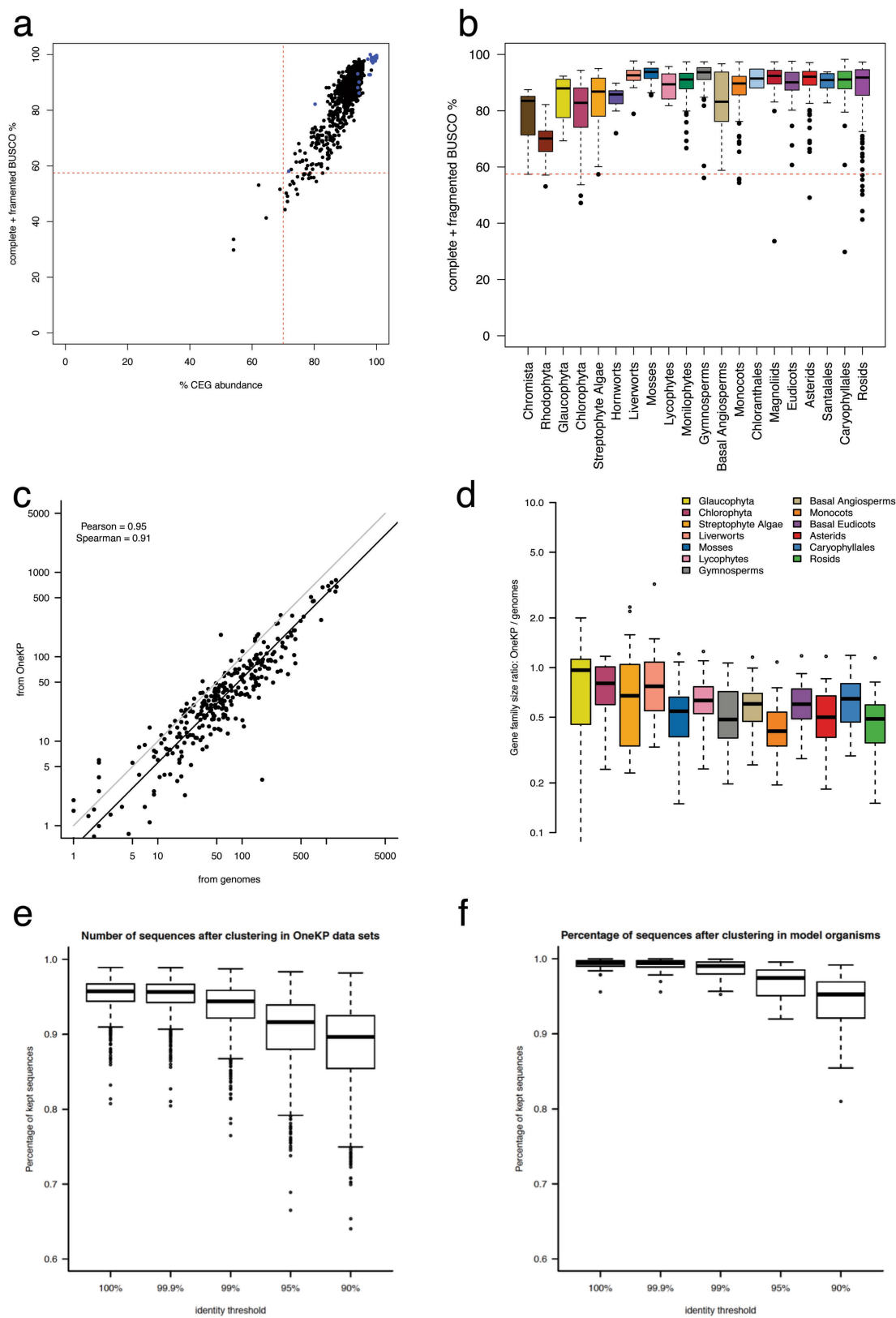
**Extended Data Fig. 1 | Mean number of MADS-box genes in the transcriptomes of different plant clades.** Type I genes are shown in green; type II genes are shown in purple and orange. Transcripts in which only a K-box was identified (which are probably partial transcripts of type II genes) are shown in orange. Data are mean  $\pm$  s.d. Dots indicate the numbers of MADS-box genes in

Clade  
individual transcriptomes. Sample sizes ( $n$ ) are as follows: liverworts,  $n = 26$ ; hornworts,  $n = 7$ ; mosses,  $n = 37$ ; lycophytes,  $n = 22$ ; eusporangiate ferns,  $n = 10$ ; leptosporangiate ferns,  $n = 62$ ; gymnosperms,  $n = 84$ ; and angiosperms,  $n = 820$ . A total of 1,068 transcriptomes were analysed for this figure.



**Extended Data Fig. 2 | RAxML phylogeny of classic type II MIKC<sup>c</sup> MADS-box genes of liverworts, mosses, lycophytes, monilophytes (ferns) and spermatophytes (seed plants).** *CgMADS1* from *Chara globularis* was used as a representative of the outgroup. Branches leading to genes from the different phyla are coloured according to the simplified phylogeny of land plants that is shown in the top left corner. The phylogenetic position of some known type II

MIKC<sup>c</sup> MADS-box genes<sup>111</sup> representative of previously described clades of MADS-box genes are indicated on the right together with the species and phylum in which these genes have been identified. The four clades of MIKC<sup>c</sup> MADS-box genes that trace back to the most recent common ancestor of Euphyllophytes are shaded in grey.



Extended Data Fig. 3 | See next page for caption.

**Extended Data Fig. 3 | Assessments of transcriptome assembly gene-family representation relative to gene-family members identified in annotated genomes.** **a**, BUSCO versus CEGMA (CEG) gene occupancy for each sample. BUSCO transcriptome completeness is given as 'complete plus fragmented' BUSCO percentage using the eukaryota\_odb9 database. CEGMA transcriptome completeness is given as conditional reciprocal best BLAST hits (see Supplementary Methods). Dotted line represents 57.5% (BUSCO) and 70% (CEGMA) gene occupancy threshold. Black dots represent IKP samples ( $n = 1,020$ ) and blue dots annotated plant genomes ( $n = 30$ ). **b**, BUSCO gene occupancy for each major clade. Boxes represent lower and upper quartiles; the black bold line represents the median and whiskers extend to the most-extreme data points. Sample sizes: Chromista,  $n = 23$ ; Rhodophyta,  $n = 18$ ; Glaucophyta,  $n = 2$ ; Chlorophyta,  $n = 94$ ; streptophyte algae,  $n = 42$ ; hornworts,  $n = 7$ ; liverworts,  $n = 18$ ; mosses,  $n = 38$ ; lycophytes,  $n = 16$ ; monilophytes,  $n = 59$ ; gymnosperms,  $n = 76$ ; ANA grade,  $n = 6$ ; monocots,  $n = 96$ ; Chloranthales,  $n = 1$ ;

magnoliids,  $n = 22$ ; CRPT grade,  $n = 29$ ; asterids,  $n = 205$ ; Caryophyllales,  $n = 48$ ; rosids,  $n = 176$ ; Saxifragales,  $n = 23$ ; Santalales,  $n = 6$ . Dotted line represents 57.5% (BUSCO) gene occupancy threshold. **c**, Scatterplot of gene-family sizes in transcriptomes versus genomes on a logarithmic scale. The grey line indicates  $x = y$ , the black line indicates a linear regression fitted to the data ( $n = 299$ ; 23 gene families in 13 species groups). Pearson and Spearman correlation coefficients ( $n = 299$ ) are indicated. **d**, Box plot of transcriptome:genome ratios of gene-family sizes for each species group. Boxes indicate upper and lower quartiles with median; whiskers extend to data points no more than  $1.5 \times$  the interquartile range ( $n = 23$ ) with outliers plotted as individual data points. **e, f**, Number of remaining sequences after filtering with cd-hit and a threshold of 100%, 99.9%, 99%, 95% or 90% in transcriptome sequences and reference genomes (Supplementary Table 8). Boxes indicate upper and lower quartiles with median; whiskers extend to data points no more than  $1.5 \times$  the interquartile range (**e**,  $n = 1,451$ ; **f**,  $n = 32$ ) with outliers plotted as individual data points.



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated
- ☒ ☐ Clearly defined error bars  
*State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on [statistics for biologists](#) may be useful.

### Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

Transcripts were assembled using SOAPdenovo-Trans assembler (version of 2012-04-05); NCBI BLAST, TransRate, CEGMA6 and BUSCO were used to assess assembly quality, translations were performed using TransPipe and Genewise 2.2.2, Gene and species tree estimates RAxML v. 8.1.17, FastTree-2 v. 2.1.5, and ExaML v. 3.0.14, ASTRAL-II v. 5.0.3 was used to estimate species trees; scripts for post-processing, DiscoVista, of trees - <https://github.com/smirarab/1kp>; genome duplications were investigated using the DupPipe, PAML, and the MAPS pipelines including the GuestTreeGen program within GenPhyloData - <https://bitbucket.org/barkerlab/maps>; analysis of gene family expansions included HMMER v3.1b2 and scripts available at <https://github.com/GrosseLab/OneKP-gene-family-evo>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data are public: Raw reads in NCBI SRA database - [http://www.onekp.com/public\\_read\\_data.html](http://www.onekp.com/public_read_data.html) ; Assembled transcripts and transcript translations - [http://www.onekp.com/public\\_data.html](http://www.onekp.com/public_data.html) ; Gene family nucleotide and amino acid fasta files - <http://jlmwiki.plantbio.uga.edu/onekp/v2/> ; Multiple sequence alignments, gene trees and species trees for single copy nuclear genes - <https://github.com/smirarab/1kp>

## Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://nature.com/authors/policies/ReportingSummary-flat.pdf)

## Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Gene and species phylogenies were estimated in order to infer: relationships across the green tree of life (Viridiplantae), the timing of genome-scale duplication events, and the timing of gene family expansions.
Research sample	RNA was isolated from young vegetative tissue from 1342 samples representing 1147 species across all major subclades of Viridiplantae, glaucophytes (Glaucophyta) and red algae (Rhodophyta) and used to generate RNA seq reads and assemblies.
Sampling strategy	Samples were collected as available in living collections. Species were chosen for RNA seq with a priority to maximize taxonomic diversity across Viridiplantae and outgroups
Data collection	RNA samples were derived from vouchered material in living collections as described in Table 1.
Timing and spatial scale	Samples were collected as available. No attempt was made to control for environmental variation
Data exclusions	RNA samples exhibiting evidence of contamination were excluded from phylogenetic analyses. Contamination was diagnosed through BLAST comparisons to ribosomal RNA and plastid gene databases.
Reproducibility	Bootstrap analyses and Bayesian posterior probabilities were estimated for all nodes in gene trees and species trees.
Randomization	Bootstrap support for nodes gene trees and species trees were estimated in a standard fashion through random resampling of columns in sequence alignments.
Blinding	No blinding was done for any of our analyses.
Did the study involve field work?	<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No

## Reporting for specific materials, systems and methods

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Unique biological materials
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Unique biological materials

---

Policy information about [availability of materials](#)

Obtaining unique materials

Most samples are available in live collections and/or herbarium vouchers.

# Dietary salt promotes cognitive impairment through tau phosphorylation

<https://doi.org/10.1038/s41586-019-1688-z>

Received: 31 May 2018

Accepted: 24 September 2019

Published online: 23 October 2019

Giuseppe Faraco<sup>1\*</sup>, Karin Hochrainer<sup>1</sup>, Steven G. Segarra<sup>1</sup>, Samantha Schaeffer<sup>1</sup>, Monica M. Santisteban<sup>1</sup>, Ajay Menon<sup>1</sup>, Hong Jiang<sup>2</sup>, David M. Holtzman<sup>2</sup>, Josef Anrather<sup>1</sup> & Costantino Iadecola<sup>1\*</sup>

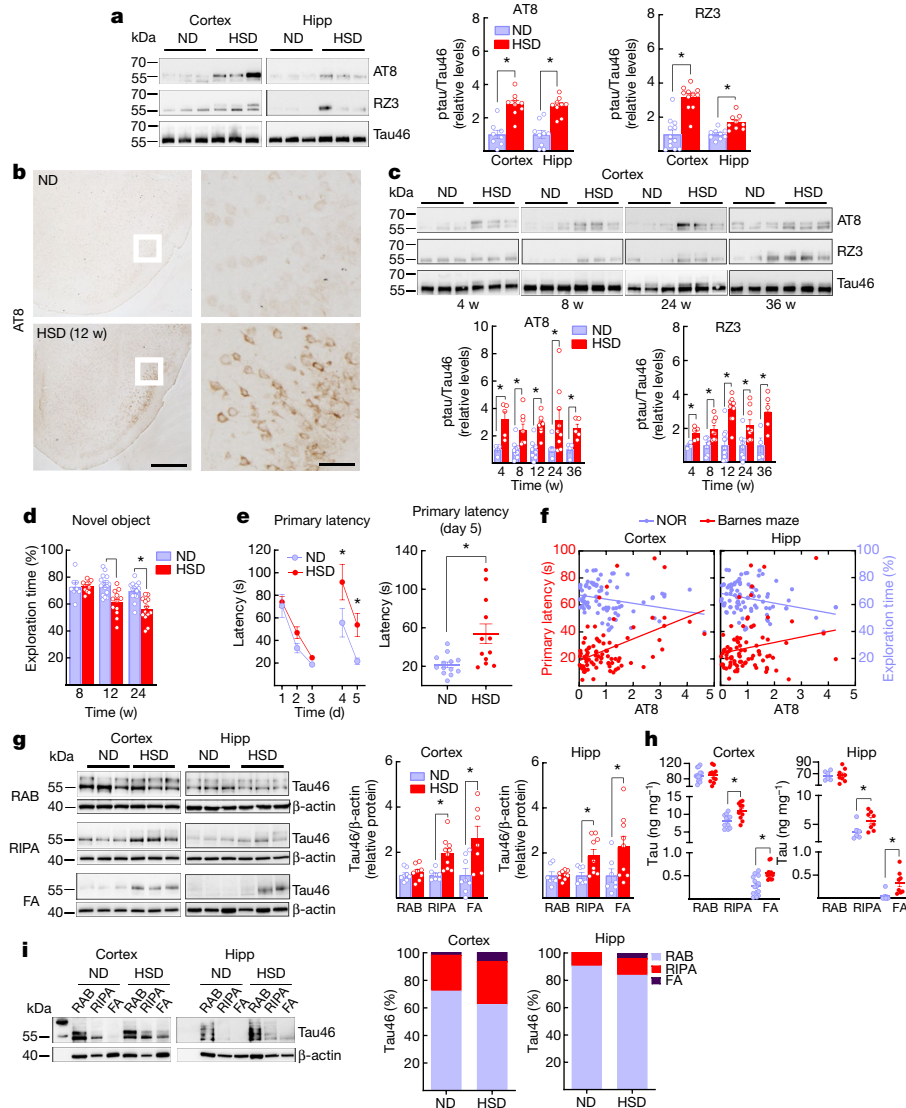
Dietary habits and vascular risk factors promote both Alzheimer's disease and cognitive impairment caused by vascular factors<sup>1–3</sup>. Furthermore, accumulation of hyperphosphorylated tau, a microtubule-associated protein and a hallmark of Alzheimer's pathology<sup>4</sup>, is also linked to vascular cognitive impairment<sup>5,6</sup>. In mice, a salt-rich diet leads to cognitive dysfunction associated with a nitric oxide deficit in cerebral endothelial cells and cerebral hypoperfusion<sup>7</sup>. Here we report that dietary salt induces hyperphosphorylation of tau followed by cognitive dysfunction in mice, and that these effects are prevented by restoring endothelial nitric oxide production. The nitric oxide deficiency reduces neuronal calpain nitrosylation and results in enzyme activation, which, in turn, leads to tau phosphorylation by activating cyclin-dependent kinase 5. Salt-induced cognitive impairment is not observed in tau-null mice or in mice treated with anti-tau antibodies, despite persistent cerebral hypoperfusion and neurovascular dysfunction. These findings identify a causal link between dietary salt, endothelial dysfunction and tau pathology, independent of haemodynamic insufficiency. Avoidance of excessive salt intake and maintenance of vascular health may help to stave off the vascular and neurodegenerative pathologies that underlie dementia in the elderly.

Vascular risk factors, including excessive salt consumption, have long been associated with cerebrovascular diseases and cognitive impairment<sup>1–3</sup>. A diet rich in salt is an independent risk factor for stroke and dementia<sup>3,8–10</sup> and has been linked to the cerebral small vessel disease that underlies vascular cognitive impairment<sup>11</sup>, a condition that is associated with endothelial dysfunction and reduced cerebral blood flow (CBF)<sup>12</sup>. In mice, a high-salt diet (HSD) induces cognitive dysfunction by targeting the cerebral microvasculature through a gut-initiated adaptive immune response mediated by T<sub>H</sub>17 lymphocytes<sup>7</sup>. The resulting increase in circulating IL-17 leads to inhibition of endothelial nitric oxide synthase (eNOS) and reduced vascular production of nitric oxide, which in turn impairs endothelial vasoactivity and lowers CBF by about 25%<sup>7</sup>. However, it remains unclear how hypoperfusion, resulting from an HSD or other vascular risk factors, leads to impaired cognition. The prevailing view is that hypoperfusion compromises the delivery of oxygen and glucose to energy-demanding brain regions that are involved in cognition<sup>12,13</sup>. However, the relatively small reduction in CBF that is associated with an HSD in mice<sup>7</sup> and vascular cognitive impairment in humans<sup>14</sup> may not be sufficient to impair cognitive function<sup>15</sup>, suggesting that vascular factors beyond cerebral perfusion are involved.

Excessive phosphorylation of the microtubule-associated protein tau promotes the formation of insoluble tau aggregates, which are thought to mediate neuronal dysfunction and cognitive impairment in Alzheimer's disease and other tauopathies<sup>16</sup>. However, accumulation of tau has also been detected in cerebrovascular pathologies

associated with endothelial dysfunction and cognitive impairment<sup>5,6</sup>. Therefore, we investigated whether tau accumulation rather than cerebral hypoperfusion contributes to the cognitive dysfunction induced by an HSD. First, we investigated whether an HSD induces phosphorylation of tau. Male C56Bl/6 mice were fed a normal diet or an HSD (4 or 8% NaCl—a commonly used model of excessive dietary salt corresponding to a 8–16-fold increase in salt content compared to regular mouse chow)<sup>7,17</sup>. Phosphorylation of tau epitopes that promote aggregation of tau and neuronal dysfunction<sup>16</sup> was assessed over time by western blotting. An HSD (8% NaCl) induced a sustained increase in phosphorylated tau (p-tau; detected using AT8 (pSer202 and pThr205) and RZ3 (pThr231) antibodies) in the neocortex and hippocampus without increasing total tau (detected using Tau 46; Fig. 1a). In the hippocampus, there was also an increase in tau phosphorylation measured using PHF13 and pSer199Ser202 antibodies (Extended Data Fig. 1a). AT8 tau phosphorylation was abolished by λ-protein phosphatase (Extended Data Fig. 1b). AT8 and RZ3 immunoreactivity were also increased in the neocortex of female mice fed an HSD (Extended Data Fig. 1c). The HSD did not increase acetylation of tau at K280, a post-translational modification that has been implicated in tau pathology<sup>18</sup> (Extended Data Fig. 1a). AT8 and MC1 immunoreactivity was detected in the piriform cortex, but we found no neurofibrillary tangles (Fig. 1b, Extended Data Fig. 1d, e). We found neither neuronal or white-matter damage, nor significant changes in astrocytes, microglia/macrophages or pericytes (Extended Data

<sup>1</sup>Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA. <sup>2</sup>Department of Neurology, Hope Center for Neurological Disorders, Knight Alzheimer's Disease Research Center, Washington University, St. Louis, MO, USA. \*e-mail: gif2004@med.cornell.edu; coi2001@med.cornell.edu



**Fig. 1 | HSD increases tau phosphorylation and insoluble tau.** **a**, HSD increases AT8 and RZ3 immunoreactivity (cortex: AT8, normal diet (ND)/HSD  $n = 8/9$ ,  $*P < 0.0001$ ; RZ3, ND/HSD  $n = 12/11$ ,  $*P < 0.0001$ ; hippocampus (Hipp): AT8, ND/HSD  $n = 9/9$ ,  $*P < 0.0001$ ; RZ3, ND/HSD  $n = 9/9$ ,  $*P = 0.0011$ ; two-tailed unpaired  $t$ -test for HSD versus ND). In all figures and legends, asterisks denote significant differences. **b**, HSD increases neuronal AT8 immunoreactivity in the piriform cortex. Right, magnified view of boxes on left (scale bars, 500  $\mu\text{m}$  (left); 100  $\mu\text{m}$  (right)). Representative images ( $n = 5$  mice per group). **c**, Time course of neocortical increase in AT8 and RZ3. AT8, 4 weeks: ND/HSD  $n = 4/5$ ,  $*P = 0.0116$ ; 8 weeks: ND/HSD  $n = 9/8$ ,  $*P = 0.0066$ ; 24 weeks: ND/HSD  $n = 8/9$ ,  $*P = 0.0152$ ; 36 weeks: ND/HSD  $n = 4/5$ ,  $*P = 0.0087$ ; RZ3, 4 weeks: ND/HSD  $n = 4/5$ ,  $*P = 0.0097$ ; 8 weeks: ND/HSD  $n = 7/8$ ,  $*P = 0.0084$ ; 24 weeks: ND/HSD  $n = 8/9$ ,  $*P = 0.0135$ ; 36 weeks: ND/HSD  $n = 4/5$ ,  $*P = 0.0204$ ; two-tailed unpaired  $t$ -test for HSD versus ND. **d**, HSD induces deficits in recognition memory. Diet:  $*P < 0.0001$ , time:  $*P = 0.0002$ ; 8 weeks: ND/HSD  $n = 8/11$ ; 12 weeks: ND/HSD  $n = 16/12$ ; 24 weeks: ND/HSD  $n = 14/13$  mice per group, two-way ANOVA and Tukey's test. **e**, HSD induces deficits in spatial memory. Diet:  $*P = 0.0048$ , time:  $*P < 0.0001$ ; ND/HSD  $n = 13/12$ , two-way repeated measures ANOVA and Bonferroni's test; primary

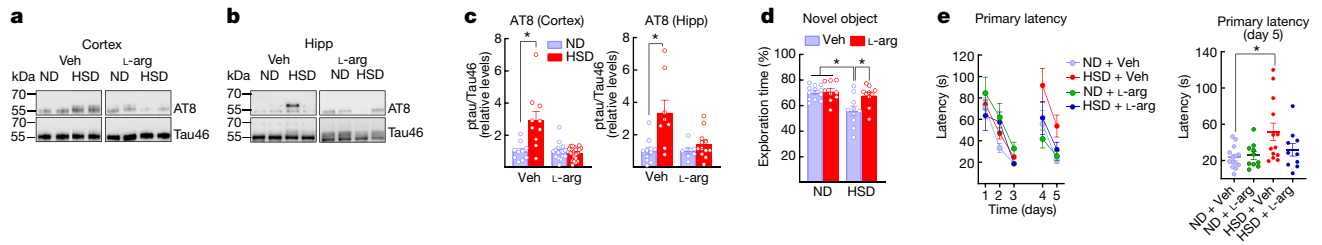
latency day 5, ND/HSD  $n = 13/12$ ,  $*P = 0.0031$  versus ND, two-tailed unpaired  $t$ -test. **f**, Neocortical and hippocampal levels of AT8 correlate with spatial learning impairment. AT8 cortex: Barnes maze  $r = 0.4491$ ,  $*P < 0.0001$ ,  $n = 84$ ; novel object recognition (NOR)  $r = -0.2621$ ,  $*P = 0.0188$ ,  $n = 80$ ; AT8 hippocampus: Barnes maze  $r = 0.2073$ ,  $*P = 0.0462$ ,  $n = 93$ ; NOR  $r = -0.2915$ ,  $*P = 0.0053$ ,  $n = 90$ ; Pearson's correlation coefficient. **g**, HSD increases levels of insoluble tau (western blotting) extracted in RIPA and FA fractions after 12 weeks. Cortex: RIPA, ND/HSD  $n = 7/10$ ,  $*P = 0.0032$ ; FA, ND/HSD  $n = 8/7$ ,  $*P = 0.0146$ ; hippocampus: RIPA, ND/HSD  $n = 7/9$ ,  $*P = 0.0418$ ; FA, ND/HSD  $n = 7/9$ ,  $*P = 0.0494$ , two-tailed unpaired  $t$ -test for HSD versus ND. **h**, HSD increases levels of insoluble tau (electrochemiluminescence). Cortex: RIPA, ND/HSD  $n = 11$ ,  $*P = 0.0050$ ; FA, ND/HSD  $n = 14/11$ ,  $*P = 0.0028$ ; hippocampus: RIPA, ND/HSD  $n = 6/8$ ,  $*P = 0.0380$ ; FA, ND/HSD  $n = 7/8$ ,  $*P = 0.0037$ ; two-tailed unpaired  $t$ -test for HSD versus ND. **i**, HSD shifts tau from the RAB fraction to the less soluble RIPA and FA fractions. Cortex: ND/HSD  $n = 9/8$ , RAB,  $P = 0.4234$ , RIPA,  $P = 0.5414$ , FA,  $*P = 0.0325$ ; hippocampus: ND/HSD  $n = 5/6$ , RAB,  $P = 0.2468$ , RIPA,  $P = 0.3290$ , FA,  $*P = 0.0152$ ; two-tailed unpaired  $t$ -test for HSD versus ND. For gel source data see Supplementary Fig. 1. Data are expressed as mean  $\pm$  s.e.m.

Fig. 2a–c). Increased AT8 immunoreactivity was also observed in the neocortex of mice fed with the 4% HSD (Extended Data Fig. 1f).

In the neocortex, AT8 immunoreactivity increased after 4 weeks and RZ3 immunoreactivity after 8 weeks of HSD, and both remained elevated for up to 36 weeks, whereas in the hippocampus AT8 immunoreactivity peaked at 12 and 36 weeks (Fig. 1c, Extended Data Fig. 1g). Starting after 12 weeks of HSD, mice exhibited difficulties in recognizing novel objects

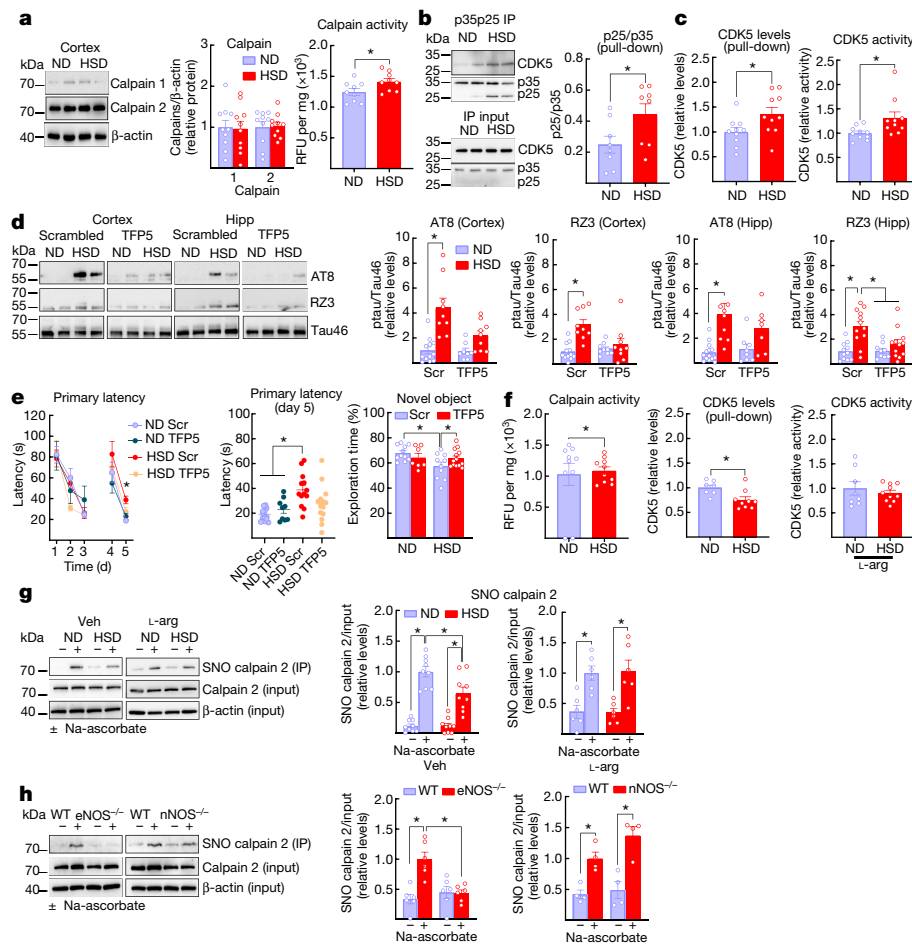
and developed a deficit in spatial memory for the Barnes maze (Fig. 1d, e, Extended Data Fig. 3a). Female mice fed an HSD also showed cognitive dysfunction (Extended Data Fig. 1c). The magnitude of phosphorylation at the AT8 and RZ3 epitopes was correlated with performance on the Barnes maze (Fig. 1f) and novel object recognition (Extended Data Fig. 3b). The HSD did not increase levels of amyloid- $\beta$  ( $\text{A}\beta_{38}$ ,  $\text{A}\beta_{40}$  or  $\text{A}\beta_{42}$ ) in the neocortex (Extended Data Fig. 3c). p-tau was also increased in





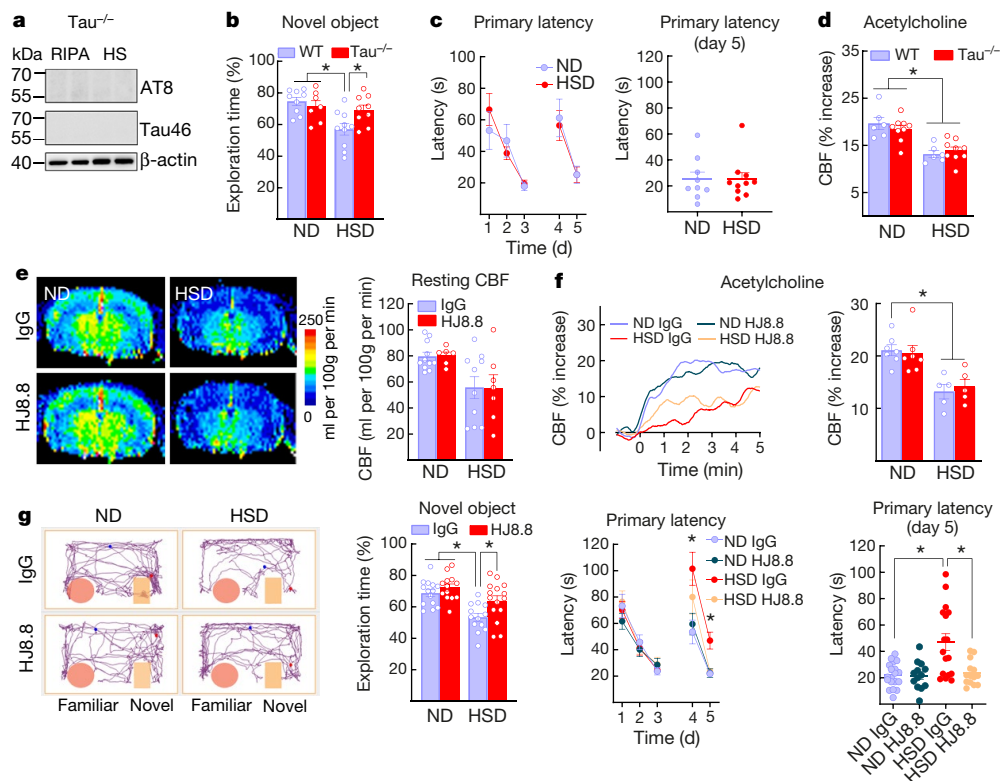
**Fig. 2 | The NO precursor L-arginine prevents the increase in p-tau induced by HSD.** **a–c**, Administration of L-arginine (10 g l<sup>-1</sup> in drinking water), starting at week 8 of HSD and continued through week 12, suppresses AT8 accumulation in neocortex (**a**) and hippocampus (**b**); quantified in **c**. Cortex: vehicle (Veh), ND/HSD  $n = 10/10$ , L-arginine (L-arg), ND/HSD  $n = 16/21$ ,  $^*P = 0.0045$ ; hippocampus: Veh, ND/HSD  $n = 10/8$ ,  $^*P = 0.0067$ , L-arg, ND/HSD  $n = 7/12$ , two-tailed unpaired  $t$ -test for HSD versus ND. **d**, **e**, L-Arginine treatment reduces the cognitive deficits

induced by HSD in both the NOR test (Veh: ND/HSD  $n = 12/10$ ; L-arg: ND/HSD  $n = 6/11$ ; diet:  $^*P = 0.0156$ , treatment:  $^*P = 0.0406$ ; two-way ANOVA and Tukey's test) and the Barnes maze (primary latency, diet:  $^*P = 0.0182$ , time:  $^*P < 0.0001$ , two-way repeated measures ANOVA and Tukey's test; primary latency day 5,  $^*P = 0.0439$ ; Kruskal–Wallis test). For gel source data see Supplementary Fig. 1. Data are expressed as mean  $\pm$  s.e.m.



**Fig. 3 | HSD induces activation of calpain and CDK5 associated with calpain denitrosylation.** **a**, HSD did not alter expression of calpain 1 or 2 (ND/HSD,  $n = 10$ ), but increased enzyme activity. ND/HSD  $n = 9/9$ ,  $^*P = 0.0404$  versus ND, two-tailed unpaired  $t$ -test. **b**, HSD increases the cleavage of p35 into p25. ND/HSD  $n = 8/8$ ,  $^*P = 0.0426$  versus ND, two-tailed unpaired  $t$ -test. **c**, HSD increases the level of CDK5 bound to p35p25 (ND/HSD  $n = 10/10$ ,  $^*P = 0.0347$ ) and CDK5 activity (ND/HSD  $n = 10/10$ ,  $^*P = 0.0274$ ; two-tailed unpaired  $t$ -test for HSD versus ND). **d**, The CDK5 peptide inhibitor TFP5 counteracts the HSD-induced increase in AT8 and RZ3 immunoreactivity. AT8, cortex: ND/HSD scrambled  $n = 10/9$ , ND/HSD TFP5  $n = 9/9$ , diet:  $^*P < 0.0001$ , treatment:  $^*P = 0.0164$ ; AT8, hippocampus: ND/HSD scrambled  $n = 11/10$ , ND/HSD TFP5  $n = 10/7$ , diet:  $^*P = 0.0004$ , treatment:  $^*P = 0.0360$ ; RZ3, cortex: ND/HSD scrambled  $n = 10/10$ , ND/HSD TFP5  $n = 10/11$ , diet:  $^*P < 0.0001$ , treatment:  $^*P = 0.0814$ ; RZ3, hippocampus: ND/HSD scrambled  $n = 12/11$ , ND/HSD TFP5  $n = 10/12$ , diet:  $^*P < 0.0001$ , treatment:  $^*P = 0.0066$ ; two-way ANOVA and Tukey's test. **e**, TFP5 rescues the spatial memory deficits induced by HSD. Primary latency, diet:

$^*P = 0.6415$ , time:  $^*P < 0.0001$ , two-way repeated measures ANOVA and Tukey's test; primary latency day 5, diet:  $^*P = 0.0016$ , treatment:  $^*P = 0.5797$ ; two-way ANOVA and Tukey's test). TFP5 also improves cognitive performance of HSD-fed mice on the NOR test. ND/HSD  $n = 6/6$ , diet:  $^*P = 0.0383$ , treatment:  $^*P = 0.1488$ ; two-way ANOVA and Tukey's test. **f**, L-Arginine counteracts the increase in calpain (ND/HSD  $n = 8/10$ ,  $^*P = 0.0335$  versus ND, two-tailed unpaired  $t$ -test) and CDK5 activity induced by HSD and reduces CDK5 bound to p35p25 (ND/HSD  $n = 7/9$ ,  $^*P = 0.0137$  versus ND, two-tailed unpaired  $t$ -test). **g**, Calpain 2 nitrosylation is reduced by HSD (ND/HSD  $n = 9/9$ , diet:  $^*P = 0.0189$ ; ascorbate:  $^*P < 0.0001$ ; two-way ANOVA and Tukey's test) and this effect is reversed by L-arginine. ND/HSD  $n = 6/6$ , diet:  $^*P = 0.9487$ , ascorbate:  $^*P < 0.0001$ , two-way ANOVA and Tukey's test. **h**, Nitrosylation is suppressed in eNOS<sup>-/-</sup> (ND/HSD  $n = 6/6$ , genotype:  $^*P = 0.0223$ , ascorbate:  $^*P = 0.0021$ , two-way ANOVA and Tukey's test), but not in nNOS<sup>-/-</sup> mice (ND/HSD  $n = 4/4$ , genotype:  $^*P = 0.0843$ , ascorbate:  $^*P < 0.0001$ ; two-way ANOVA and Tukey's test). For gel source data see Supplementary Fig. 1. Data are expressed as mean  $\pm$  s.e.m.



**Fig. 4 | HSD-induced cognitive dysfunction is not observed in  $\tau^{-/-}$  mice and is prevented by tau antibodies despite cerebrovascular insufficiency.** **a**, AT8 and Tau46 are absent in  $\tau^{-/-}$  mice in RIPA and heat-stable (HS) RIPA fractions. Representative blots from  $n = 5$  ND-fed  $\tau^{-/-}$  mice. **b**, **c**, HSD does not alter cognition in  $\tau^{-/-}$  mice on the NOR test (wild-type (WT): ND/HSD  $n = 9/10$ ;  $\tau^{-/-}$ : ND/HSD  $n = 7/9$ ; diet:  $P = 0.0055$ , genotype:  $P = 0.1827$ , two-way ANOVA and Tukey's test) or the Barnes maze (ND/HSD  $n = 9/10$ , diet:  $P = 0.9348$ , time:  $P < 0.0001$ , two-way ANOVA and Tukey's test). **d**, The increase in CBF produced by neocortical application of acetylcholine is reduced in  $\tau^{-/-}$  mice (ND/HSD wild-type,  $n = 6/6$ ,  $\tau^{-/-}$ ,  $n = 9/9$ ; diet:  $P < 0.0001$ , genotype:  $P = 0.7920$ , two-way ANOVA and Tukey's test). **e**, Anti-tau antibodies (HJ8.8, 50 mg per kg per week, intraperitoneal injection) do not prevent the reduction in resting CBF induced by HSD (IgG: ND/HSD  $n = 11/9$ ; HJ8.8: ND/HSD  $n = 6/5$ ; diet:  $P = 0.0061$ ,

treatment:  $P = 0.9367$ , two-way ANOVA and Tukey's test). **f**, HJ8.8 does not rescue the CBF response to acetylcholine (IgG: ND/HSD  $n = 5/5$ ; HJ8.8: ND/HSD  $n = 5/5$ ; diet:  $P = 0.0005$ , treatment:  $P = 0.8516$ , two-way ANOVA and Tukey's test). **g**, HJ8.8 ameliorates the cognitive dysfunction induced by HSD on both the NOR test (IgG: ND/HSD  $n = 15/13$ ; HJ8.8: ND/HSD  $n = 13/15$ ; diet:  $P = 0.0001$ , treatment:  $P = 0.0054$ , two-way ANOVA and Tukey's test) and the Barnes maze test (primary latency: IgG: ND/HSD  $n = 19/15$ ; HJ8.8: ND/HSD  $n = 13/14$ ; time:  $P < 0.0001$ , diet:  $P = 0.0358$ , two-way repeated measures ANOVA and Tukey's test; primary latency day 5: IgG: ND/HSD  $n = 19/16$ ; HJ8.8: ND/HSD  $n = 13/14$ ;  $P = 0.0202$ , Kruskal–Wallis and Dunn's test). Images on left show representative tracks. For gel source data see Supplementary Fig. 1. Data are expressed as mean  $\pm$  s.e.m.

mice with arterial hypertension induced by angiotensin-II (Extended Data Fig. 3d, e) or in a model of A $\beta$  accumulation (Tg2576 mice; Extended Data Fig. 3f), both of which are associated with endothelial dysfunction and cognitive impairment<sup>19,20</sup>.

To determine whether an HSD alters the solubility of tau—a key determinant of its harmful effects<sup>16</sup>—we measured tau levels in brain tissue after sequential biochemical extraction in RAB (salt buffer), RIPA (detergent buffer) or 70% formic acid (FA), which contain, respectively, soluble, less soluble and highly insoluble tau. In samples taken from mice after 12 weeks of the HSD, tau in the RIPA and FA fractions was increased over that in samples from mice fed a normal diet, reflecting an increase in insoluble tau (Fig. 1g–i). Hypothermia, which does not cause cognitive impairment, also increased p-tau<sup>21</sup>, but, unlike the HSD, did not lead to an increase in insoluble species (Extended Data Fig. 3g, h). These observations indicate that an HSD not only promotes hyperphosphorylation of tau, but also its aggregation.

As the nitric oxide precursor L-arginine counteracts the endothelial nitric oxide deficit in mice fed an HSD<sup>7</sup>, mice were given L-arginine in their drinking water (10 g l<sup>-1</sup>) during the last 4 weeks of the 12-week HSD course. L-Arginine suppressed accumulation of p-tau and prevented HSD-induced cognitive dysfunction (Fig. 2a–e, Extended Data Fig. 4a), without affecting the HSD-induced increase in circulating IL-17

(Extended Data Fig. 4b). Consistent with a key role of endothelial nitric oxide deficiency in accumulation of p-tau, eNOS-null (*Nos3*<sup>-/-</sup>) mice fed a normal diet showed elevated p-tau that was not increased further by an HSD (Extended Data Fig. 4c, d).

Cyclin-dependent kinase 5 (CDK5), a kinase that is responsible for tau hyperphosphorylation<sup>16</sup>, is tightly regulated by its binding partner p35<sup>22,23</sup>. Calpains cleave p35 bound to CDK5 into p25, which results in dysregulated activation of CDK5 and hyperphosphorylation of tau<sup>22,23</sup>. As reduced endothelial nitric oxide may lead to tau phosphorylation by activating CDK5 via p25<sup>24</sup>, we investigated whether the HSD influenced the activity of calpain and CDK5. Calpain 2 is more abundant than calpain 1 in the neocortex (Fig. 3a) and is colocalized with CDK5 in neurons (Extended Data Fig. 4e, f). Mice fed an HSD showed increased calpain activity compared to mice fed a normal diet, and this led to an increase in the p25/p35 ratio and activation of CDK5 (Fig. 3a–c). Other CDK5 substrates besides tau, such as DARPP-32<sup>25</sup>, were not phosphorylated (Extended Data Fig. 4g). Administration of the CDK5 peptide inhibitor TFP5 (40 mg kg<sup>-1</sup> twice per week; intraperitoneal)<sup>26</sup> attenuated phosphorylation of tau and prevented cognitive dysfunction (Fig. 3d, e), without blunting the HSD-induced increase in circulating IL-17 (Extended Data Fig. 4h). L-Arginine prevented the HSD-induced activation of calpain and CDK5 (Fig. 3f) but did not alter calpain levels (Extended Data Fig. 4i).

The HSD did not increase the activity of GSK3 $\beta$ , an enzyme that has been implicated in tau phosphorylation<sup>16</sup>, or the expression of the prolyl *cis/trans* isomerase PIN-1, which regulates tau dephosphorylation<sup>27</sup> (Extended Data Fig. 5a, b).

When calpain has been activated by Ca<sup>2+</sup> it is regulated mainly by its endogenous inhibitor calpastatin and by nitrosylation by nitric oxide<sup>28</sup>, which suppresses its activity<sup>29</sup>. As the HSD did not reduce calpastatin expression (Extended Data Fig. 5c), we used the biotin switch assay to investigate the effect of an HSD on calpain nitrosylation. Calpain nitrosylation was reduced in mice fed an HSD compared with mice fed a normal diet, and this effect was reversed by L-arginine (Fig. 3g). Nitrosylation was markedly suppressed in eNOS-null mice, but not in nNOS-null (*Nos1<sup>-/-</sup>*) mice, attesting to the key role of eNOS-derived nitric oxide in regulating nitrosylation and activity of calpain (Fig. 3h). As the HSD reduced CDK5 nitrosylation (Extended Data Fig. 5d), it is unlikely that this modification, which activates the enzyme<sup>30</sup>, is involved in the effects of the HSD.

Finally, we used tau-null (*Mapt<sup>-/-</sup>*) mice and anti-tau antibodies to examine the relative contributions of p-tau and neurovascular dysfunction to the cognitive deficits induced by an HSD. Tau-null mice that were fed an HSD for 12 weeks did not exhibit cognitive impairment, but still showed a marked attenuation of the increase in CBF evoked by neocortical application of acetylcholine (Fig. 4a–d), a response mediated by eNOS-derived nitric oxide<sup>31</sup>. Similarly, wild-type mice treated with anti-tau antibodies (HJ8.8)<sup>32</sup> or control IgG (50 mg per kg per week; intraperitoneal) for the last 4 weeks of the 12-week HSD regimen showed improved cognitive function (Fig. 4g) compared with untreated mice, despite having reduced resting CBF and an attenuated CBF response to acetylcholine (Fig. 4e, f). HJ8.8 lowered p-tau in the hippocampus (measured by AT8 immunoreactivity; Extended Data Fig. 5e) but did not blunt the HSD-induced increase in circulating IL-17 (Extended Data Fig. 5f). The HSD did not affect the increase in CBF induced by neural activity in either tau-null mice or mice treated with HJ8.8 (Extended Data Fig. 5g).

These observations indicate that the cognitive dysfunction associated with HSD is mediated by a deficit in endothelial nitric oxide that results from denitrosylation of calpain. This deficit leads to activation of CDK5 and phosphorylation of tau in neurons (Extended Data Fig. 5h). Notably, the hypoperfusion and neurovascular dysfunction that also result from the endothelial nitric oxide deficit do not mediate the cognitive impairment. Rather, other aspects of endothelial function are involved—namely, endothelial nitric oxide maintaining calpain homeostasis and preventing CDK5 dysregulation and tau hyperphosphorylation.

Although the HSD used in mice, in which the salt content is 8–16-fold higher than in normal mouse chow<sup>7,17</sup>, may exceed the highest reported levels of human salt consumption (12.5–20 g per day or 3–5 times the recommended level of 4–5 g per day)<sup>33</sup>, our data provide a previously unrecognized link between dietary habits, vascular dysfunction and tau pathology, independent of cerebral hypoperfusion. Such relationships may play a role in the frequent coexistence of vascular and neurodegenerative pathologies in conditions that cause dementia, such as Alzheimer's disease and frontotemporal dementia<sup>12,13</sup>. Whereas the avoidance of excessive salt consumption may help to prevent tau pathology, therapeutic efforts to counteract cerebrovascular dysfunction need to go beyond rescuing cerebral perfusion and target vascular mediators that govern the neurovascular interactions that are essential for cognitive health.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1688-z>.

- Scarmeas, N., Anastasiou, C. A. & Yannakoulia, M. Nutrition and prevention of cognitive impairment. *Lancet Neurol.* **17**, 1006–1015 (2018).
- van de Rest, O., Berendsen, A. A., Haveman-Nies, A. & de Groot, L. C. Dietary patterns, cognitive decline, and dementia: a systematic review. *Adv. Nutr.* **6**, 154–168 (2015).
- Kendig, M. D. & Morris, M. J. Reviewing the effects of dietary salt on cognition: mechanisms and future directions. *Asia Pac. J. Clin. Nutr.* **28**, 6–14 (2019).
- De Strooper, B. & Karran, E. The cellular phase of Alzheimer's disease. *Cell* **164**, 603–615 (2016).
- Nation, D. A. et al. Pulse pressure in relation to tau-mediated neurodegeneration, cerebral amyloidosis, and progression to dementia in very old adults. *JAMA Neurol.* **72**, 546–553 (2015).
- Kim, H. J. et al. Assessment of extent and role of tau in subcortical vascular cognitive impairment using 18F-AV1451 positron emission tomography imaging. *JAMA Neurol.* **75**, 999–1007 (2018).
- Faraco, G. et al. Dietary salt promotes neurovascular and cognitive dysfunction through a gut-initiated TH17 response. *Nat. Neurosci.* **21**, 240–249 (2018).
- Fiocco, A. J. et al. Sodium intake and physical activity impact cognitive maintenance in older adults: the NuAge Study. *Neurobiol. Aging* **33**, 829.e821–829.e28, (2012).
- Gardener, H., Rundek, T., Wright, C. B., Elkind, M. S. & Sacco, R. L. Dietary sodium and risk of stroke in the Northern Manhattan study. *Stroke* **43**, 1200–1205 (2012).
- Blumenthal, J. A. et al. Lifestyle and neurocognition in older adults with cognitive impairments: a randomized trial. *Neurology* **92**, e212–e223 (2019).
- Heye, A. K. et al. Blood pressure and sodium: association with MRI markers in cerebral small vessel disease. *J. Cereb. Blood Flow Metab.* **36**, 264–274 (2016).
- Iadecola, C. The pathobiology of vascular dementia. *Neuron* **80**, 844–866 (2013).
- Sweeney, M. D. et al. Vascular dysfunction—the disregarded partner of Alzheimer's disease. *Alzheimers Dement.* **15**, 158–167 (2019).
- Shi, Y. et al. Cerebral blood flow in small vessel disease: a systematic review and meta-analysis. *J. Cereb. Blood Flow Metab.* **36**, 1653–1667 (2016).
- Marshall, R. S. et al. Recovery of brain function during induced cerebral hypoperfusion. *Brain* **124**, 1208–1217 (2001).
- Wang, Y. & Mandelkow, E. Tau in physiology and pathology. *Nat. Rev. Neurosci.* **17**, 5–21 (2016).
- Zhao, Y. et al. Sodium intake regulates glucose homeostasis through the PPAR $\delta$ /adiponectin-mediated SGLT2 pathway. *Cell Metab.* **23**, 699–711 (2016).
- Min, S. W. et al. Critical role of acetylation in tau-mediated neurodegeneration and cognitive deficits. *Nat. Med.* **21**, 1154–1162 (2015).
- Iadecola, C. et al. SOD1 rescues cerebral endothelial dysfunction in mice overexpressing amyloid precursor protein. *Nat. Neurosci.* **2**, 157–161 (1999).
- Faraco, G. et al. Perivascular macrophages mediate the neurovascular and cognitive dysfunction associated with hypertension. *J. Clin. Invest.* **126**, 4674–4689 (2016).
- Arendt, T., Stieler, J. T. & Holzer, M. Tau and tauopathies. *Brain Res. Bull.* **126**, 238–292 (2016).
- Lee, M. S. et al. Neurotoxicity induces cleavage of p35 to p25 by calpain. *Nature* **405**, 360–364 (2000).
- Patrick, G. N. et al. Conversion of p35 to p25 deregulates Cdk5 activity and promotes neurodegeneration. *Nature* **402**, 615–622 (1999).
- Austin, S. A. & Katusic, Z. S. Loss of endothelial nitric oxide synthase promotes p25 generation and tau phosphorylation in a murine model of Alzheimer's disease. *Circ. Res.* **119**, 1128–1134 (2016).
- Bibb, J. A. et al. Phosphorylation of DARPP-32 by Cdk5 modulates dopamine signalling in neurons. *Nature* **402**, 669–671 (1999).
- Shukla, V. et al. A truncated peptide from p35, a Cdk5 activator, prevents Alzheimer's disease phenotypes in model mice. *FASEB J.* **27**, 174–186 (2013).
- Kimura, T. et al. Isomerase Pin1 stimulates dephosphorylation of tau protein at cyclin-dependent kinase (Cdk5)-dependent Alzheimer phosphorylation sites. *J. Biol. Chem.* **288**, 7968–7977 (2013).
- Ono, Y., Saido, T. C. & Sorimachi, H. Calpain research for drug discovery: challenges and potential. *Nat. Rev. Drug Discov.* **15**, 854–876 (2016).
- Etwebi, Z., Landesberg, G., Preston, K., Eguchi, S. & Scalia, R. Mechanistic role of the calcium-dependent protease calpain in the endothelial dysfunction induced by MPO (myeloperoxidase). *Hypertension* **71**, 761–770 (2018).
- Qu, J. et al. S-Nitrosylation activates Cdk5 and contributes to synaptic spine loss induced by  $\beta$ -amyloid peptide. *Proc. Natl. Acad. Sci. USA* **108**, 14330–14335 (2011).
- Iadecola, C. The neurovascular unit coming of age: a journey through neurovascular coupling in health and disease. *Neuron* **96**, 17–42 (2017).
- Yanamandra, K. et al. Anti-tau antibodies that block tau aggregate seeding *in vitro* markedly decrease pathology and improve cognition *in vivo*. *Neuron* **80**, 402–414 (2013).
- Powles, J. et al. Global, regional and national sodium intakes in 1990 and 2010: a systematic analysis of 24 h urinary sodium excretion and dietary surveys worldwide. *BMJ Open* **3**, e003733 (2013).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

Most of the methods used in this study are well established in the laboratory and have been described in detail in previous publications<sup>7,20,34</sup>. Here we provide only a brief description.

### Mice

All procedures were approved by the institutional animal care and use committee of Weill Cornell Medicine (animal protocol number: 0807-777A). Studies were conducted, according to the ARRIVE guidelines (<https://www.nc3rs.org.uk/arrive-guidelines>), in the following lines of mice: C57BL/6 (JAX), B6.129X1-Maptm1Hnd (Tau<sup>-/-</sup>, JAX, Stock #007251), B6.129P2-Nos3<sup>tm1Unc</sup>/J (eNOS<sup>-/-</sup>, JAX, Stock #002684), B6.129S4-Nos1<sup>tm1Plh</sup>/J (nNOS<sup>-/-</sup>, JAX, Stock #002986), 129S6.Cg-Tg(APP<sup>SWE</sup>)2576Kha N20+? (Taconic, Stock #279) and Tg(Camk2a-tTA)*IMmay Fgf14*<sup>Tg(tetO-MAPT<sup>P301L</sup>J4510Kha</sup>/J (rTg4510, JAX, Stock #024854). Unless otherwise indicated, 8-week-old male mice were used.

### High-salt diet

Male or female mice (8 weeks old) received normal chow (0.5% NaCl) and tap water ad libitum (normal diet (ND)) or sodium-rich chow (4 or 8% NaCl) and tap water containing 1% NaCl ad libitum (HSD) for 4–36 weeks as stated<sup>7</sup>.

### In vivo treatments

The nitric oxide precursor L-arginine (10 g/l; Sigma) was administered in the drinking water starting after 8 weeks of HSD and continuing until 12 weeks. ND- and HSD-fed mice were treated (intraperitoneally (i.p.), weekly) with 50 mg/kg of anti-tau (HJ8.8) or mouse IgG1 isotype control (Clone MOPC-21; bioXcell) antibodies for the last 4 weeks of the HSD treatment period (12 weeks) before behavioural and cerebrovascular studies. HJ8.8 is a high-affinity antibody generated against human tau that can recognize both human and mouse tau ( $K_d$  (dissociation constant) = 0.926 nM to mouse tau). In other experiments, ND- and HSD-fed mice were treated (i.p. twice a week) with 40 mg/kg of TPF5 (KEAFWDRCLSVINLMSSKMLQINAYARAARRAARR) or scrambled peptide (GGGFWDRLSGKGMSSKGGGINAYARAARRAARR) (Peptide 2.0)<sup>35</sup> for the last 4 weeks of the HSD treatment period (12 weeks) before behavioural and molecular studies.

### General surgical procedures for CBF studies

Mice were anaesthetized with isoflurane (induction, 5%; maintenance, 2%). The trachea was intubated and mice were artificially ventilated with a mixture of N<sub>2</sub> and O<sub>2</sub>. One of the femoral arteries was cannulated for recording mean arterial pressure (MAP) and collecting blood samples for blood gas analysis<sup>36</sup>. Rectal temperature was maintained at 37 °C. End tidal CO<sub>2</sub>, monitored by a CO<sub>2</sub> analyser (Capstar-100, CWE Inc.), was maintained at 2.6–2.7% to provide a pCO<sub>2</sub> of 30–40 mm Hg and a pH of 7.3–7.437. After surgery, isoflurane was discontinued and anaesthesia was maintained with urethane (750 mg/kg, i.p.) and chloralose (50 mg/kg, i.p.). Throughout the experiment the level of anaesthesia was monitored by testing motor responses to a tail pinch.

### Monitoring CBF

A small craniotomy (2 × 2 mm) was performed to expose the parietal cortex, the dura was removed and the site was superfused with Ringer's solution (37 °C; pH 7.3–7.4)<sup>20</sup>. CBF was continuously monitored at the site of superfusion with a laser-Doppler probe (Perimed) positioned stereotactically ~0.5 mm above the cortical surface and connected to a data acquisition system (PowerLab). CBF values are expressed as percentage increases relative to the resting level.

### Protocol for CBF experiments

After MAP and blood gases stabilized, CBF responses were recorded<sup>7</sup>. The whisker-barrel cortex was activated for 60 s by stroking the contralateral vibrissae, and the evoked changes in CBF were recorded.

The endothelium-dependent vasodilator acetylcholine (ACh; 100 μM, Sigma) was superfused onto the exposed neocortex for 5 min and the associated changes in CBF were recorded by laser-Doppler flowmetry. CBF and MAP data were collected using Chart 5 Pro (v.5.5.6).

### Measurement of resting CBF by ASL-MRI

CBF was assessed quantitatively using arterial spin labelling magnetic resonance imaging (ASL-MRI), performed on a 7.0-Tesla 70/30 Bruker Biospec small-animal MRI system with 450 mT/m gradient amplitude and a 4,500 T/m/s slew rate. A volume coil was used for transmission and a surface coil for reception. Anatomical localizer images were acquired to find the transversal slice approximately corresponding to bregma + 0.5 mm. This position was used for subsequent ASL-MRI, which was based on a flow-sensitive alternating inversion recovery rapid acquisition with relaxation enhancement (FAIR-RARE) pulse sequence labelling the inflowing blood by global inversion of the equilibrium magnetization. One axial slice was acquired with a field of view of 15 × 15 mm, spatial resolution of 0.117 × 0.117 × 1 mm, TE (echo time) of 5.368 ms, effective TE of 48.32 ms, recovery time of 10 s and a RARE (rapid imaging with refocused echoes) factor of 72. Twenty-two turbo inversion recovery values ranging from 30 to 2,300 ms were used, and the inversion slab thickness was 4 mm. For computation of resting CBF (rCBF), the Bruker ASL perfusion processing macro was used. It uses a published model<sup>37</sup> and includes steps to mask out the background. The masked rCBF images were exported to Analyze format on the MRI console. The ASL images were analysed by ImageJ and the average CBF value is reported as ml per 100 g of tissue per minute<sup>7</sup>.

### Osmotic minipump implantation for delivery of angiotensin-II

Osmotic minipumps containing vehicle (saline) or angiotensin-II (600 ng/kg/min) were implanted subcutaneously under isoflurane anaesthesia. Systolic blood pressure was monitored in awake mice using tail-cuff plethysmography<sup>20</sup>. Forty-two days later, mice were anaesthetized and their brains were collected for assessment of tau phosphorylation.

### Hypothermia

C57BL/6 mice (12 weeks old) were anaesthetized by injection of ketamine/xylazine (100/10 mg/kg). Rectal temperature was continuously monitored and kept at 37 °C (normothermia) or 30 °C (hypothermia) using a thermostatically controlled heating pad. Mice were killed 30 min after anaesthesia and their brains were collected and frozen on dry ice. Tissues were kept at –80 °C until processing for immunoblot analysis.

### Immunoblot analysis

Cortex (~80–90 mg) and hippocampus (~15 mg) isolated from ND- and HSD-fed mice were sonicated in 800 and 500 μl of RIPA buffer, respectively (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.5% deoxycholic acid, 0.1% SDS, 1 mM EDTA pH 8.0, 1% IGEPAL CA-630, 1 mM Na<sub>3</sub>VO<sub>4</sub>, 20 mM NaF and one tablet per 10 ml of cOmplete, EDTA-free Protease Inhibitor Cocktail, Millipore Sigma) and equal volumes were mixed with SDS sample buffer, boiled and analysed on 10% or 10–20% Novex WedgeWell gels (Thermo Fisher Scientific). Proteins were transferred to PVDF membranes (Millipore), blocked at room temperature for 1 h with 5% milk in TBS, and incubated overnight at 4 °C, with primary antibodies (see Reporting Summary) in 5% BSA in TBS/0.1% Tween-20 (TBST). Membranes were washed in TBST and incubated with goat anti-mouse or rabbit secondary antibodies conjugated to horseradish peroxidase (Santa Cruz Biotechnology) for 1 h at room temperature, and protein bands were visualized with Clarity Western ECL Substrate (Bio Rad) on a Bio Rad ChemiDoc MP Imaging System. Quantification was performed using Image Laboratory v.6.0 (Bio Rad).

### Preparation of heat-stable RIPA fractions for tau enrichment

After homogenization in cold RIPA buffer and centrifugation, 150 μl of the supernatant containing the proteins was boiled at 100 °C for 10 min.

# Article

Samples were cooled on ice for 20 min and then centrifuged at 20,000g at 4 °C for 15 min. The supernatant corresponding to the heat-stable fraction was then harvested. This method is used to isolate proteins resistant to heat, including tau and other microtubule-associated proteins. Thus, endogenous immunoglobulins are precipitated during the boiling process and eliminated from the supernatant. The proteins were then mixed with equal volumes of SDS sample buffer, boiled and analysed on 10% Novex WedgeWell gels (Thermo Fisher Scientific). Although tau protein is partially lost during the boiling process, the heat-stable samples are enriched with tau (Extended Data Fig. 5h). Furthermore, boiling substantially improves the specificity of certain antibodies, such as AT8, RZ3 or MC1<sup>38</sup>.

## Tau dephosphorylation

After overnight dialysis to remove phosphatase inhibitors, protein samples (40 µl) were incubated with 5 µl of 10× NEBuffer for Protein MetalloPhosphatases (PMP), 5 µl of 10 mM MnCl<sub>2</sub> and 1 µl of Lambda protein phosphatase (Lambda PP, New England Biolabs) at 30 °C for 3 h. Reactions were stopped by addition of SDS sample buffer and boiling for 5 min at 100 °C.

## Brain tissue protein extraction

Extraction was performed as described previously<sup>32</sup>. The cortex (~80–90 mg) and hippocampus (~15 mg) of each brain were homogenized by sonication in 800 and 300 µl, respectively, of RAB buffer (100 mM MES, 1 mM EDTA, 0.5 mM MgSO<sub>4</sub>, 750 mM NaCl, 20 mM NaF, 1 mM Na<sub>3</sub>VO<sub>4</sub>, supplemented by EDTA-free Protease Inhibitor Cocktail, Millipore Sigma). In brief, the samples were centrifuged at 50,000g for 20 min at 4 °C using an Optima MAX-TLA 120.2 Ultracentrifuge (Beckman). The supernatants were collected as RAB-soluble fractions and pellets were resuspended in identical volumes of RIPA buffer (150 mM NaCl, 50 mM Tris, 0.5% deoxycholic acid, 1% Triton X-100, 0.5% SDS, 25 mM EDTA, pH 8.0, 20 mM NaF, 1 mM Na<sub>3</sub>VO<sub>4</sub> supplemented by EDTA-free Protease Inhibitor Cocktail, Millipore Sigma), and centrifuged at 50,000g for 20 min at 4 °C. The supernatants were collected as RIPA-soluble fractions. The pellets were sonicated in 70% FA (300 µl for cortex and 125 µl for hippocampus), and centrifuged at 50,000g for 20 min at 4 °C. The supernatants were collected as 70% FA fractions. All fractions were stored at –80 °C until analysed. For western blotting, an aliquot of 100 µl of the FA fractions was evaporated in a Savant SpeedVac concentrator at 45 °C for 1 h. The samples were resuspended in 100 µl SDS sample buffer with the addition of 1 µl of 10 N NaOH, sonicated and then boiled for 5 min.

## Measurement of tau, Aβ and IL-17

Tau, Aβ and IL-17 were measured using an electrochemiluminescence-based multi-array method through the Quickplex SQ 120 system (Meso Scale Diagnostics LLC). Tau and Aβ peptide levels (Aβ<sub>38</sub>, Aβ<sub>40</sub> and Aβ<sub>42</sub>) were measured in the RAB, RIPA and FA fractions of both cortex and hippocampus using the MSD Mouse Total Tau (K151DSD) and MSD V-PLEX Aβ Peptide Panel 1 (4G8) (K15199) kits according to the manufacturer's protocol. IL-17A was measured in the serum of ND- and HSD-fed mice using the MSD V-PLEX Mouse IL-17A Kit (K152RFD) according to the manufacturer's protocol.

## Immunohistochemistry

After 12 weeks of ND or HSD, mice were anaesthetized with intraperitoneal pentobarbital (200 mg/kg), and then perfused transcardially with cold PBS, followed by cold 4% paraformaldehyde (PFA) in PBS. The brains were removed and immersed first in 4% PFA overnight and then in 70% ethanol for 3 days. Brains were then embedded in paraffin and cut into 6-µm sections using a microtome. After rehydration and antigen retrieval in preheated citrate buffer (10 µM) for 30 min, brain sections were immersed in 3% H<sub>2</sub>O<sub>2</sub> and then blocked with 100% Sniper (Biocare Medical) for 1 h. After blocking, sections were incubated for 2.5 days at 4 °C with the AT8 and MC1 antibodies (1:250 and 1:100 in

1:50 Sniper in PBS, respectively) and thereafter processed for 1 h with the biotinylated secondary antibody in 1% normal donkey serum PBS (anti-mouse IgG1, Jackson ImmunoResearch). Reactions were visualized with the ABC-complex (Vectorlabs) and 3,3'-diaminobenzidine. A Nikon light microscope was used to visualize the signal associated with each antibody.

## Immunofluorescence

After 12 weeks of ND or HSD, mice were anaesthetized with intraperitoneal pentobarbital (200 mg/kg), and then perfused transcardially with cold PBS, followed by cold 4% paraformaldehyde (PFA) in PBS. The brains were removed and immersed first in 4% PFA overnight. Sections (thickness, 40 µm) were cut through the whole brain using a vibratome. After blocking with 5% normal donkey serum in 0.5% Triton-X/PBS, sections were incubated over the weekend at 4 °C with antibodies against NEUN (1:200, mouse, Millipore Sigma, MAB377), GFAP (1:200, mouse, Millipore Sigma, G3893), IBA1 (1:200, rabbit, Wako, 019-19741), CD13 (1:200, goat, R&D Systems, AF2335), calpain 2 (1:100, rabbit, Santa Cruz, sc-373966) or CDK5 (1:100, rabbit, Santa Cruz, sc-6247) in 0.5% Triton-X in PBS and thereafter processed for 2 h with FITC, Cy3 or Cy5 secondary antibodies in 0.5% Triton-X in PBS. An epifluorescence microscope (IX83 Inverted Microscope, Olympus) or a confocal microscope (Leica TCS SP5) was used to visualize the signal associated with each antibody.

## Fluoro-Jade B staining

After rehydration, brain sections (6 µm) were immersed in 1% sodium hydroxide in 80% alcohol for 5 min, followed by 2 min in 70% alcohol and 2 min in distilled water. The slides were transferred to a solution of 0.06% potassium permanganate for 10 min on a shaker. After rinsing for 2 min in distilled water, the slides were immersed in the staining solution (0.0004% Fluoro-Jade B, Millipore Sigma, in 0.1% acetic acid) for 20 min in the dark. Finally, the slides were rinsed three times for 1 min in distilled water and then placed on a slide warmer, set at approximately 50 °C, until they were fully dry. The dry slides were cleared by immersion in xylene for at least a minute before coverslipping with DPX (Sigma).

## TUNEL staining

After rehydration and antigen retrieval according to the manufacturer's protocol (In situ Cell Death Detection Kit, Fluorescein, Roche, #11 684 795 910), brain sections were blocked for 30 min in Tris-HCl, 0.1 M pH 7.5, containing 3% BSA and 20% normal donkey serum and then incubated with the TUNEL reaction mixture for 60 min at 37 °C in a humidified atmosphere in the dark. After washing, slides were evaluated by an epifluorescence microscope (IX83 Inverted Microscope, Olympus). A positive control was obtained by pre-incubating brain slices with DNase I recombinant (3,000 U/ml in 50 mM Tris-HCl, pH 7.5, 1 mg/ml BSA), for 10 min at room temperature, to induce DNA strand breaks.

## Klüver–Barrera white-matter staining

The Klüver–Barrera stain was performed using the Luxol Fast Blue Stain Kit (ScyTek Laboratory Inc.). Brains were removed after transcardiac perfusion with PBS and 4% PFA and sectioned with a vibratome (thickness, 40 µm), and the positive (blue-stained) area in the corpus callosum was quantified by ImageJ.

## Thioflavin S staining

After mounting on slides and post-fixation with 4% PFA in PBS for 10 min, coronal brain sections (40 µm) were washed and labelled with 0.05% (w/v) thioflavine-S in 50% (v/v) ethanol for 10 min as previously described<sup>39</sup>. An epifluorescence microscope (IX83 Inverted Microscope, Olympus) was used to visualize the FITC signal associated with thioflavine-S.

## Calpain activity

Calpain activity was measured using a Calpain Activity Assay Kit from AbCam<sup>40,41</sup>. In brief, fresh cortex and hippocampus were homogenized



in the extraction buffer provided with the kit, which specifically extracts cytosolic proteins without contamination of cell membrane or lysosome proteases and prevents auto-activation of calpain during the extraction procedure. The fluorometric assay is based on the detection of cleavage of the calpain substrate Ac-LLY-AFC. Ac-LLY-AFC emits blue light ( $\lambda_{\text{max}} = 400 \text{ nm}$ ); upon cleavage of the substrate by calpain, free AFC emits a yellow-green fluorescence ( $\lambda_{\text{max}} = 505 \text{ nm}$ ), which can be quantified using a fluorometer or a fluorescence plate reader. The specificity of the signal was confirmed using the calpain inhibitor Z-LLY-FMK (100–200  $\mu\text{M}$ ). The activity is expressed as relative fluorescent units (RFU) per milligram of protein for each sample.

#### p35/p25 and GSK3 $\beta$ immunoprecipitation

Immunoprecipitation was performed with anti-p35p25 (Cell Signaling), anti-GSK3 $\beta$  (Cell Signaling) or anti-rabbit monoclonal IgG1 isotype control antibodies (Santa Cruz Biotechnology). Samples were incubated overnight with the primary antibodies and then with protein-A sepharose (p35p25) (GE Healthcare Life Sciences) or protein-G Dynabeads (GSK3 $\beta$ ) (Thermo Fisher Scientific) for 2 h at 4 °C. Precipitates were used for measurement of CDK5 or GSK3 $\beta$  activity. Immunoprecipitation was confirmed by loading the samples onto 10% Tris-glycine SDS polyacrylamide gels and western blotting as described above.

#### Detection of S-nitrosylation with the biotin-switch technique

S-nitrosylated calpain 2 was detected using the biotin-switch technique, as previously described<sup>42</sup>. In brief, samples were sonicated in 800  $\mu\text{l}$  of RIPA buffer containing 0.1 mM neocuproine and, after centrifugation, protein concentrations were measured. Cysteine thiol groups in 1 mg of proteins were blocked with 10% S-methylmethane thiosulfonate (MMTS) (Sigma). After protein-precipitation with 100% acetone, sodium ascorbate was added to the sample to convert each S-nitrothiol (SNO) to a free thiol via a transnitrosation reaction to generate O-nitrosoascorbate. Next, each nascent free thiol (previously an SNO site) was biotinylated with biotin-HPDP (Pierce). Biotinylated proteins were then pulled down using avidin beads and analysed on 10% Novex WedgeWell gels (Thermo Fisher Scientific). Before avidin pull-down, a small fraction of each sample was collected to determine protein 'input'. The degree of pull-down correlates with protein S-nitrosylation of calpain 2 or CDK5, which was detected with an antibody against the two proteins. Nitrosylation of calpain 2 or CDK5 is expressed as the ratio between the pull-down signal and the input corrected for  $\beta$ -actin levels.

#### CDK5 and GSK3 $\beta$ activity

CDK5 activity in brain lysates was determined after pull-down with p25/p35 antibody (Cell Signaling) from 500  $\mu\text{g}$  total protein using a synthetic histone H1 peptide substrate (PKTPKKAKKL, Enzo Life Sciences). GSK3 $\beta$  activity was determined after pull-down with GSK3 $\beta$  antibody (Cell Signaling) from 100  $\mu\text{g}$  total protein using phospho-glycogen synthase peptide-2a as substrate (Tocris). Phosphorylation reactions were initiated by mixing bead-coupled CDK5 with 40  $\mu\text{l}$  reaction buffer containing the following: 50 mM HEPES.KOH (pH 7.4), 5 mM MgCl<sub>2</sub>, 0.05% BSA, 50  $\mu\text{M}$  substrate, 50  $\mu\text{M}$  cold ATP, 1 mM dithiothreitol, 1 $\times$  complete protease inhibitors without EDTA (Roche Applied Biosciences) and 5 Ci/mmol <sup>32</sup>P-ATP. Companion reactions for every sample were executed in the presence of the CDK5 inhibitor (R)-CR8, Tocris) (10  $\mu\text{M}$ ) or the GSK3 $\beta$  inhibitor (CHIR 99021, Tocris) (10  $\mu\text{M}$ ) to correct for non-specific activity. Reactions were incubated at 30 °C for 30 min, after which they were terminated by spotting on P81 phosphocellulose cation-exchange chromatography paper. Filters were washed four times for 2 min in 0.5% phosphoric acid, and the remaining radioactivity was quantified in a scintillation counter by the Cherenkov method.

#### Novel object recognition test

The NOR test was conducted under dim light in a plastic box. Stimuli consisted of plastic objects that varied in colour and shape but had

similar sizes<sup>43,44</sup>. A video camera mounted on the wall directly above the box was used to record the testing session for off-line analysis. Mice were acclimated to the testing room and chamber for one day before testing. Twenty-four hours after habituation, mice were placed in the same box in the presence of two identical sample objects and were allowed to explore for 5 min. After an inter-session interval of 24 h, mice were placed in the same box but one of the two objects was replaced by a novel object. Mice were allowed to explore for 5 min. Exploratory behaviour was later assessed manually by an experimenter blinded to the treatment group. Exploration of an object was defined as the mouse sniffing the object or touching the object while looking at it. Placing the forepaws on the objects was considered as exploratory behaviour but climbing on the objects was not. A minimal exploration time for both objects (total exploration time) during the test phase (~5 s) was used. The amount of time taken to explore the novel object was expressed as a percentage of the total exploration time and provides an index of recognition memory<sup>43,44</sup>. Any-Maze v5.3 was used for collection and analysis of the behavioural data.

#### The Barnes maze test

The Barnes maze consisted of a circular open surface (90 cm in diameter) elevated to 90 cm on four wooden legs<sup>45</sup>. There were 20 circular holes (5 cm in diameter) equally spaced around the perimeter, and positioned 2.5 cm from the edge of the maze. No wall and no intra-maze visual cues were placed around the edge. A wooden plastic escape box (11  $\times$  6  $\times$  5 cm) was positioned beneath one of the holes. Two neon lamps and a buzzer were used as aversive stimuli. The Any-Maze tracking system (Stoelting) was used to record the movement of mice on the maze. Extra-maze visual cues consisted of objects within the room (table, computer, sink, door and so on) and the experimenter. Mice were tested in groups of seven to ten, and between trials they were placed into cages, which were placed in a dark room adjacent to the test room for the inter-trial interval (20–30 min). No habituation trial was performed. The acquisition phase consisted of three consecutive training days with three trials per day with the escape hole located at the same location across trials and days. On each trial a mouse was placed into a start tube located in the centre of the maze, the start tube was raised, and the buzzer was turned on until the mouse entered the escape hole. After each trial, mice remained in the escape box for 60 s before being returned to their cage. Between trials the maze floor was cleaned with 10% ethanol in water to minimize olfactory cues. For each trial mice were given 3 min to locate the escape hole, after which they were guided to the escape hole or placed directly into the escape box if they failed to enter the escape hole. Four parameters of learning performance were recorded: (1) the latency to locate (primary latency) and (2) enter the escape hole (total latency), (3) the number of errors made and (4) the distance travelled before locating the escape hole<sup>45</sup>. When a mouse dipped its head into a hole that did not provide escape, it was considered an error. On days 4 and 5, the location of the escape hole was moved 180° from its previous location (reverse learning) and two trials per day were performed. Any-Maze v.5.3 was used for collection and analysis of the behavioural data.

#### Statistics

Sample size was determined using power analysis based on work previously published by our laboratory on the effects of dietary salt on CBF regulation and cognitive function<sup>7</sup>. On these bases, 10–15 mice per group were required in studies involving assessment of cognitive function and cerebrovascular function<sup>7,20</sup>. Mice were randomized to the different experimental conditions and treatments using the random number generator function (RANDBETWEEN) in Microsoft Excel. Analysis of the data was performed in a blinded fashion using GraphPad Prism (v.7.0). All data were tested for normal distribution using the Shapiro–Wilk test. Intergroup differences were analysed using the two-tailed unpaired *t*-test for single comparison or using one- or two-way ANOVA (with Tukey's or

# Article

Bonferroni's post-hoc analysis) for multiple comparisons. Non-normally distributed data were tested using the Mann–Whitney *U* test for single comparison or the Kruskal–Wallis test for multiple comparisons. Data are expressed as mean  $\pm$  s.e.m. and differences are considered statistically significant at  $P < 0.05$ .

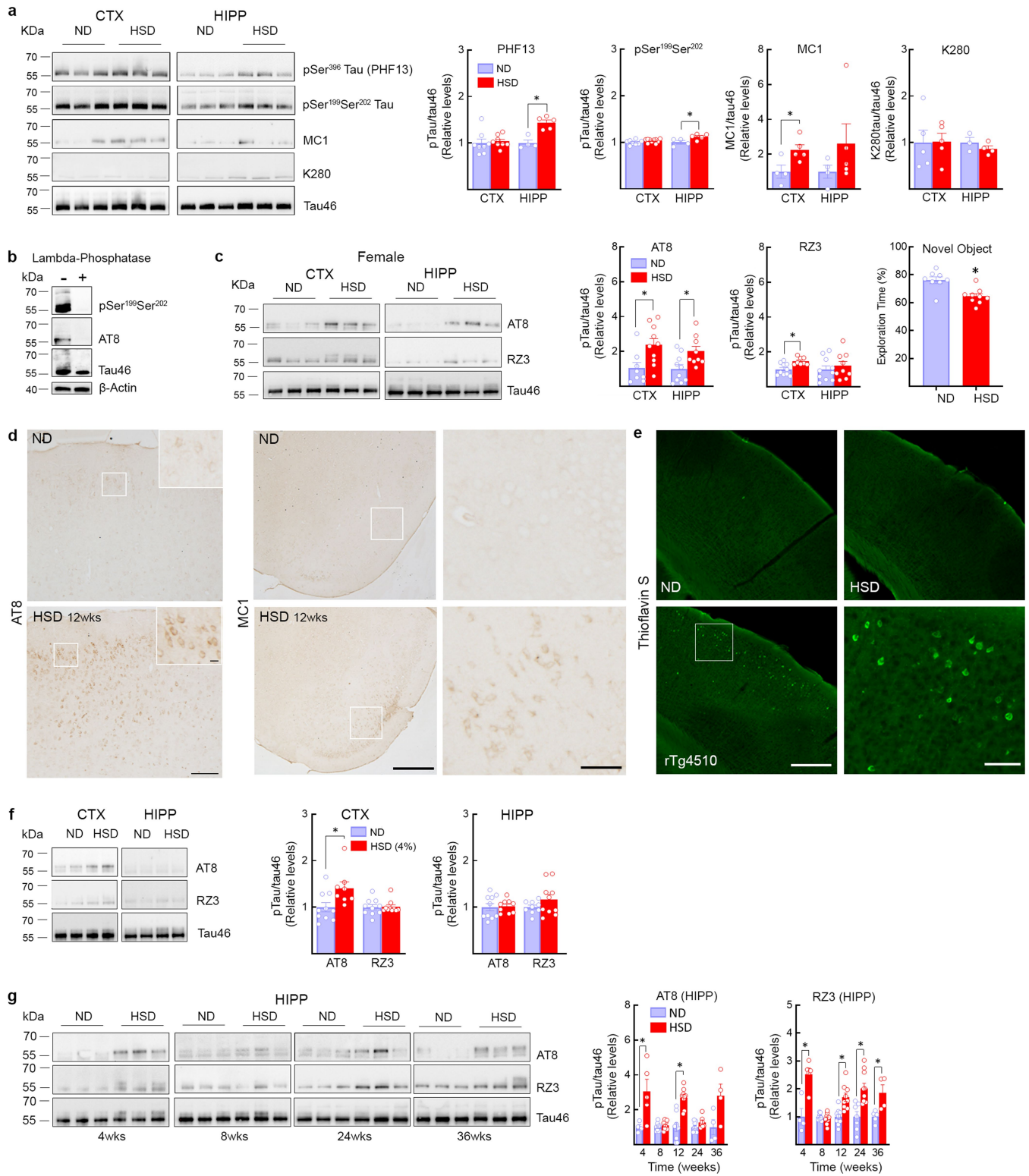
## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

Source data include final quantifications from in vivo animal work.

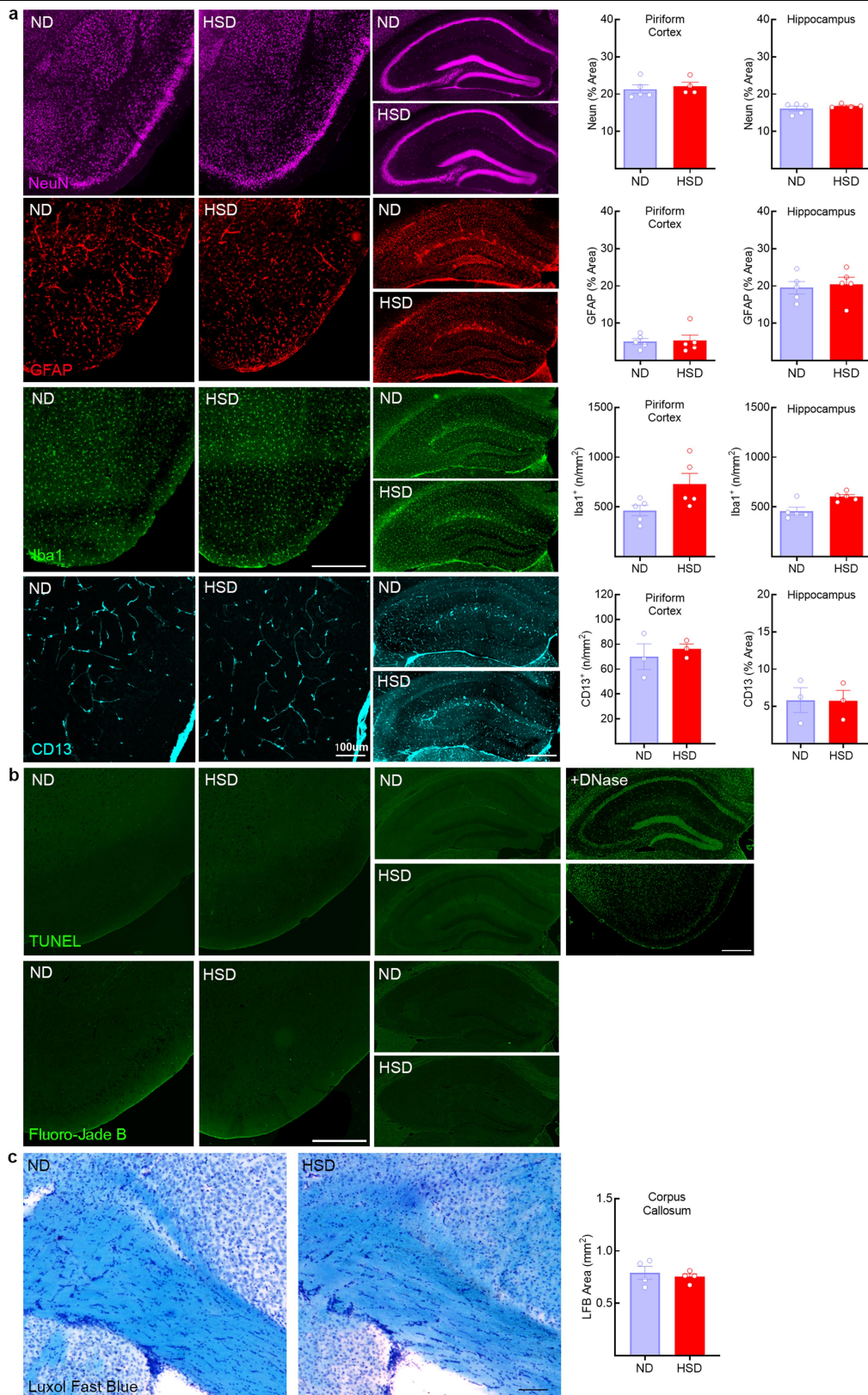
34. Hochrainer, K. et al. The ubiquitin ligase HERC3 attenuates NF- $\kappa$ B-dependent transcription independently of its enzymatic activity by delivering the RelA subunit for degradation. *Nucleic Acids Res.* **43**, 9889–9904 (2015).
  35. Shukla, V. et al. TFP5, a peptide inhibitor of aberrant and hyperactive Cdk5/p25, attenuates pathological phenotypes and restores synaptic function in CK-p25Tg mice. *J. Alzheimers Dis.* **56**, 335–349 (2017).
  36. Faraco, G. et al. Circulating endothelin-1 alters critical mechanisms regulating cerebral microcirculation. *Hypertension* **62**, 759–766 (2013).
  37. Kober, F. et al. High-resolution myocardial perfusion mapping in small animals in vivo by spin-labeling gradient-echo imaging. *Magn. Reson. Med.* **51**, 62–67 (2004).
  38. Petry, F. R. et al. Specificity of anti-tau antibodies when analyzing mice models of Alzheimer's disease: problems and solutions. *PLoS One* **9**, e94251 (2014).
  39. Faraco, G. et al. Hypertension enhances A $\beta$ -induced neurovascular dysfunction, promotes  $\beta$ -secretase activity, and leads to amyloidogenic processing of APP. *J. Cereb. Blood Flow Metab.* **36**, 241–252 (2016).
  40. Voit, A. et al. Reducing sarcolipin expression mitigates Duchenne muscular dystrophy and associated cardiomyopathy in mice. *Nat. Commun.* **8**, 1068 (2017).
  41. Liu, W. et al. Metabolic stress-induced cardiomyopathy is caused by mitochondrial dysfunction due to attenuated Erk5 signaling. *Nat. Commun.* **8**, 494 (2017).
  42. Forrester, M. T., Foster, M. W., Benhar, M. & Stamler, J. S. Detection of protein S-nitrosylation with the biotin-switch technique. *Free Radic. Biol. Med.* **46**, 119–126 (2009).
  43. Cohen, S. J. & Stackman, R. W. Jr Assessing rodent hippocampal involvement in the novel object recognition task. A review. *Behav. Brain Res.* **285**, 105–117 (2015).
  44. Grayson, B. et al. Assessment of disease-related cognitive impairments using the novel object recognition (NOR) task in rodents. *Behav. Brain Res.* **285**, 176–193 (2015).
  45. O'Leary, T. P. & Brown, R. E. Optimization of apparatus design and behavioral measures for the assessment of visuo-spatial learning and memory of mice on the Barnes maze. *Learn. Mem.* **20**, 85–96 (2013).
- Acknowledgements** We thank P. Davies for providing the RZ3, MC1 and PHF1 antibodies and Y. Li for sharing the Quickplex SQ 120 system (Meso Scale Diagnostics LLC). This study was supported by National Institutes of Health grants R37-NS089323 (C.I.) and 1R01-NS095441 (C.I.), by a grant from the Cure Alzheimer's Fund (G.F. and C.I.) and by a Scientist Development Grant from the American Heart Association (G.F.). Support from the Feil Family Foundation is gratefully acknowledged.
- Author contributions** G.F. performed western blotting experiments, behavioural tests and cerebrovascular studies, and analysed data. K.H. performed experiments on CDK5 and GSK3 $\beta$  activity and analysed data. S.G.S. performed western blotting experiments, behavioural tests and immunohistochemistry. S.S. and M.M.S. performed experiments on the effects of hypertension on tau. A.M. performed immunohistochemistry experiments. H.J. and D.M.H. provided the HJ8.8 antibody. J.A. supervised the molecular aspects of the study and edited the manuscript. G.F. and C.I. designed and supervised the entire study and wrote the manuscript.
- Competing interests** D.M.H. is listed as an inventor on a patent licensed by Washington University to C2N Diagnostics and subsequently AbbVie on the therapeutic use of anti-tau antibodies; co-founded and is on the scientific advisory board of C2N Diagnostics; and is on the scientific advisory board of Denali, Genentech, and Proclara. C.I. is on the scientific advisory board of Broadview Ventures.
- Additional information**
- Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1688-z>.
- Correspondence and requests for materials** should be addressed to G.F. or C.I.
- Peer review information** *Nature* thanks Nikolaos Scarmeas, Berislav Zlokovic and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
- Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1** | See next page for caption.

**Extended Data Fig. 1 | HSD (4 or 8%) induced tau phosphorylation: brain localization, sex differences and time course.** **a**, HSD (8% NaCl) increases tau phosphorylation on Ser396 (PHF13) and on Ser199Ser202 in the hippocampus (HIPP) but not in the neocortex (CTX) (HIPP: PHF13, ND/HSD  $n = 4/5$ ,  $*P = 0.0016$ ; Ser<sup>199</sup>Ser<sup>202</sup>, ND/HSD  $n = 4/5$ ,  $*P = 0.0337$ ; two-tailed unpaired  $t$ -test versus ND), whereas acetylation of tau on Lys280 (K280) is not affected. MC1 immunoreactivity increases in both neocortex and hippocampus in HSD-fed mice but reaches statistical significance only in neocortex (MC1, ND/HSD  $n = 4/5$ ,  $*P = 0.0321$  versus ND, two-tailed unpaired  $t$ -test). **b**, Tau phosphorylation on Ser199Ser202 and Ser202Thr205 is abolished after treatment of brain samples with lambda phosphatase. **c**, HSD increases AT8 immunoreactivity (left graph) in neocortex and hippocampus of female mice, but RZ3 (middle) increases only in neocortex (AT8, cortex: ND/HSD  $n = 8/10$ ,  $*P = 0.0159$ ; hippocampus: ND/HSD  $n = 10/9$ ,  $*P = 0.0151$ ; RZ3, cortex: ND/HSD  $n = 10/8$ ,  $*P = 0.0117$ ; two-tailed unpaired  $t$ -test for HSD versus ND). Right, HSD induces a deficit in NOR in female mice (ND/HSD  $n = 8/9$ ,  $*P = 0.0017$  versus ND, two-tailed unpaired  $t$ -test). **d**, HSD increases AT8 immunoreactivity in neuronal cell bodies of the somatosensory

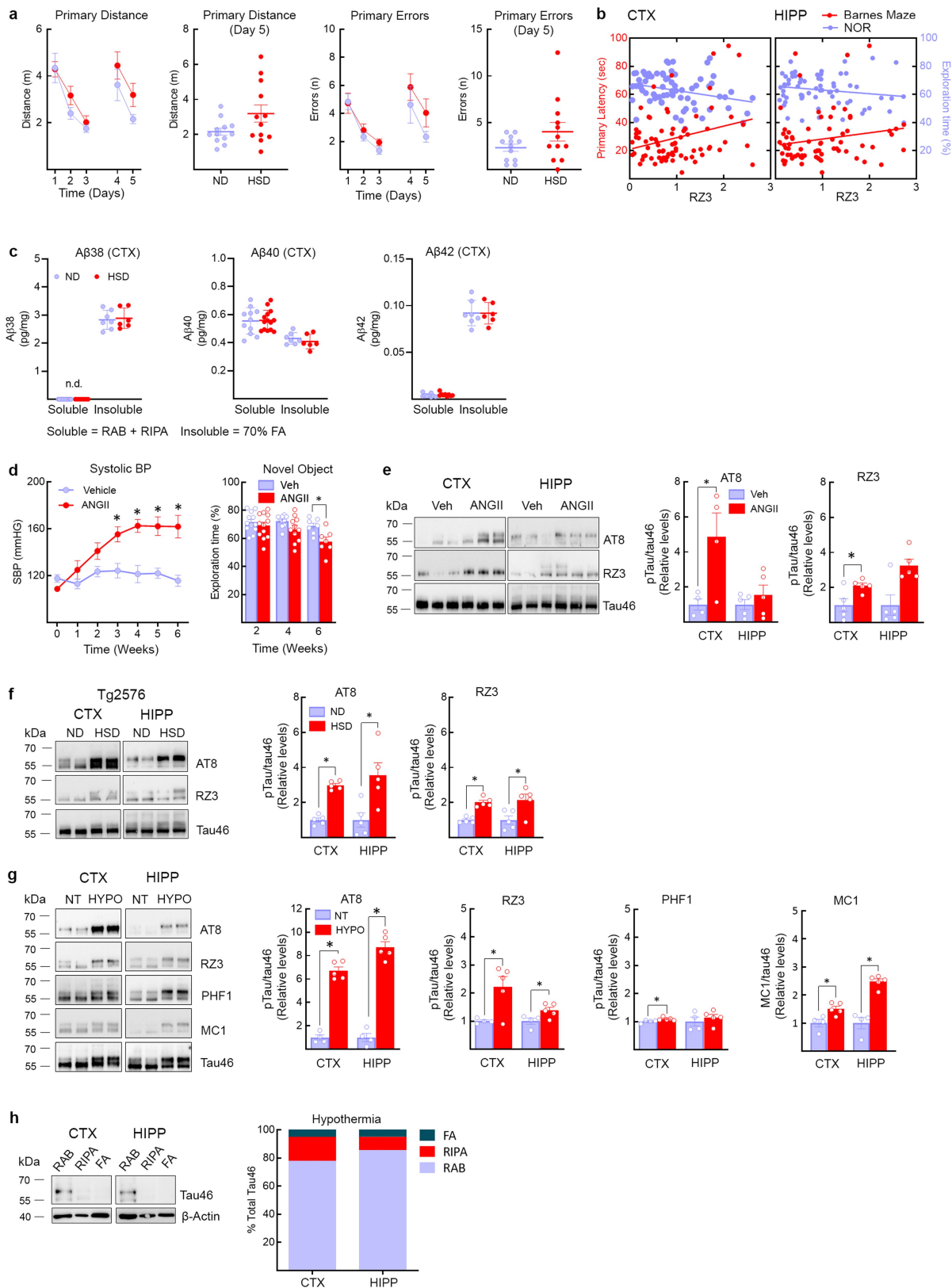
cortex (scale bars, 100  $\mu\text{m}$  (main images); 10  $\mu\text{m}$  (insets)) and MC1 immunoreactivity in neuronal bodies of the pyriform cortex (scale bars, 500  $\mu\text{m}$  (main images); 100  $\mu\text{m}$  (insets)). Representative images from ND- and HSD-fed mice ( $n = 5$  per group). **e**, Thioflavin S staining is not present in mice fed an HSD, indicating absence of neurofibrillary tangles, which can be observed in rTg4510 mice (scale bars, 500  $\mu\text{m}$  (main images); 100  $\mu\text{m}$  (inset)). Representative images from  $n = 5$  ND- and HSD-fed mice and  $n = 3$  rTg4510 mice. **f**, HSD (4%) increases AT8 immunoreactivity in the neocortex but not in the hippocampus (AT8, cortex: ND/HSD  $n = 5/5$ ,  $*P = 0.0148$  versus ND, two-tailed unpaired  $t$ -test). RZ3 was not increased in both regions. **g**, Time course of the increase in AT8 and RZ3 induced by HSD in the hippocampus. AT8 levels are increased after 4 weeks of HSD. RZ3 levels are increased after 4, 12, 24 and 36 weeks of HSD (AT8, 4 weeks: ND/HSD  $n = 4/5$ ,  $*P = 0.0386$ ; 12 weeks: ND/HSD  $n = 9/9$ ,  $*P < 0.0001$ ; RZ3, 4 weeks: ND/HSD  $n = 4/5$ ,  $*P = 0.0041$ ; RZ3, 12 weeks: ND/HSD  $n = 9/9$ ,  $*P = 0.0011$ ; 24 weeks: ND/HSD  $n = 7/10$ ,  $*P = 0.0017$ ; 36 weeks: ND/HSD  $n = 5/4$ ,  $*P = 0.0188$ ; two-tailed unpaired  $t$ -test for HSD versus ND). For gel source data see Supplementary Fig. 1. Data are expressed as mean  $\pm$  s.e.m.



**Extended Data Fig. 2 | Effect of HSD on neurons, astrocytes, microglia/macrophages, pericytes and white-matter integrity. a,** HSD (NaCl 8%) does not affect neurons (NEUN), astrocytes (GFAP), microglia/macrophages (IBA-1) (cortex: microglia, ND/HSD  $n = 5/5$ ,  $P = 0.0570$ ; hippocampus: ND/HSD  $n = 5/5$ ,  $P = 0.0556$ ; two-tailed unpaired  $t$ -test for HSD versus ND) or pericytes (CD13) in both the piriform cortex and the hippocampus (scale bars, 500  $\mu\text{m}$  (except where specified)). **b,** No evidence of neuronal cell death is observed in HSD-fed

mice by Fluoro-Jade B or TUNEL staining (scale bar, 500  $\mu\text{m}$ ). +DNase indicates positive control for TUNEL staining. Representative images from ND- and HSD-fed mice ( $n = 5$  per group). **c,** Klüver-Barrera stain shows no white-matter damage in the corpus callosum of HSD-fed mice (scale bar, 100  $\mu\text{m}$ ). Representative images from ND- and HSD-fed mice ( $n = 4$  per group). Data are expressed as mean  $\pm$  s.e.m.

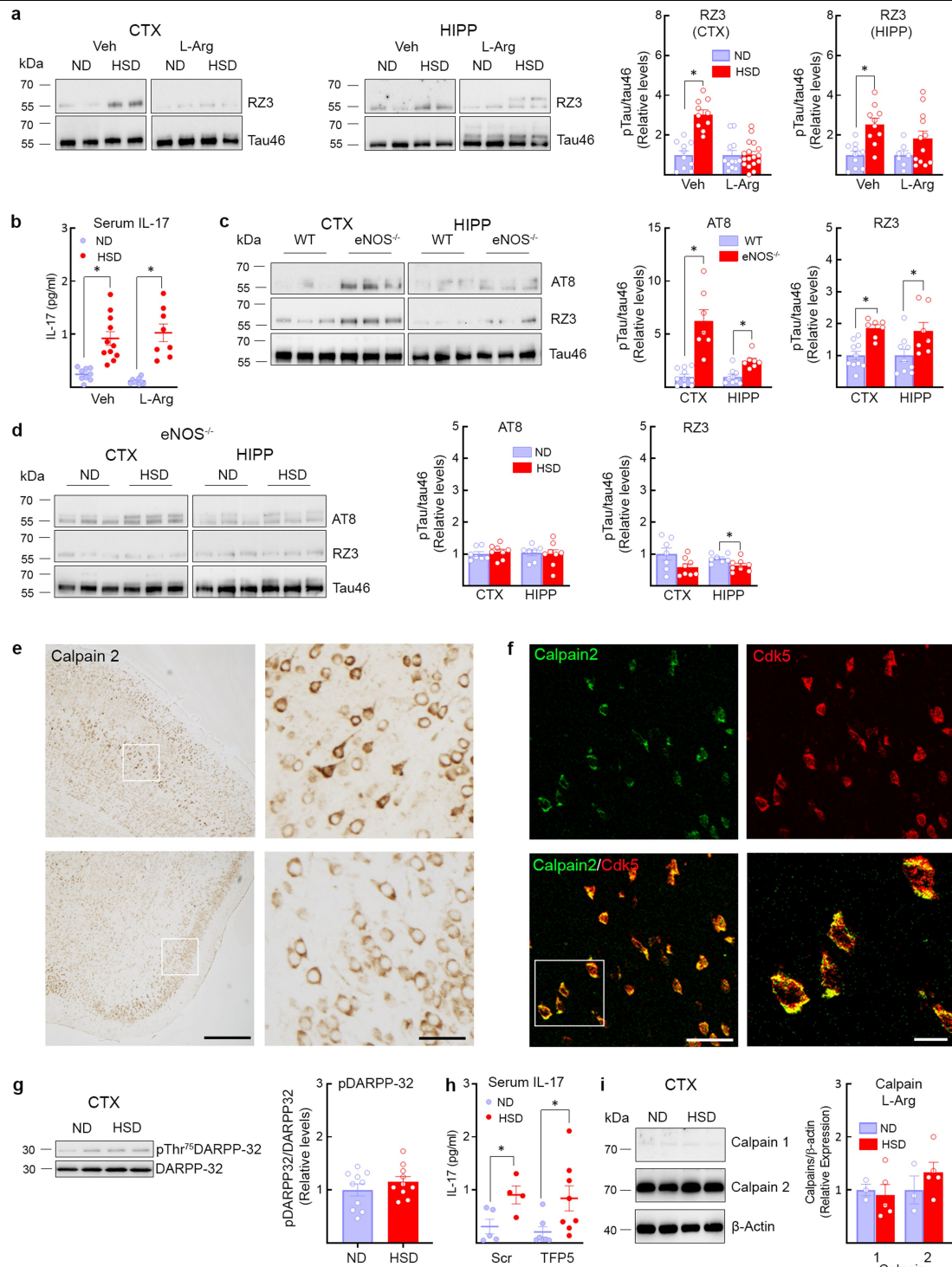




Extended Data Fig. 3 | See next page for caption.

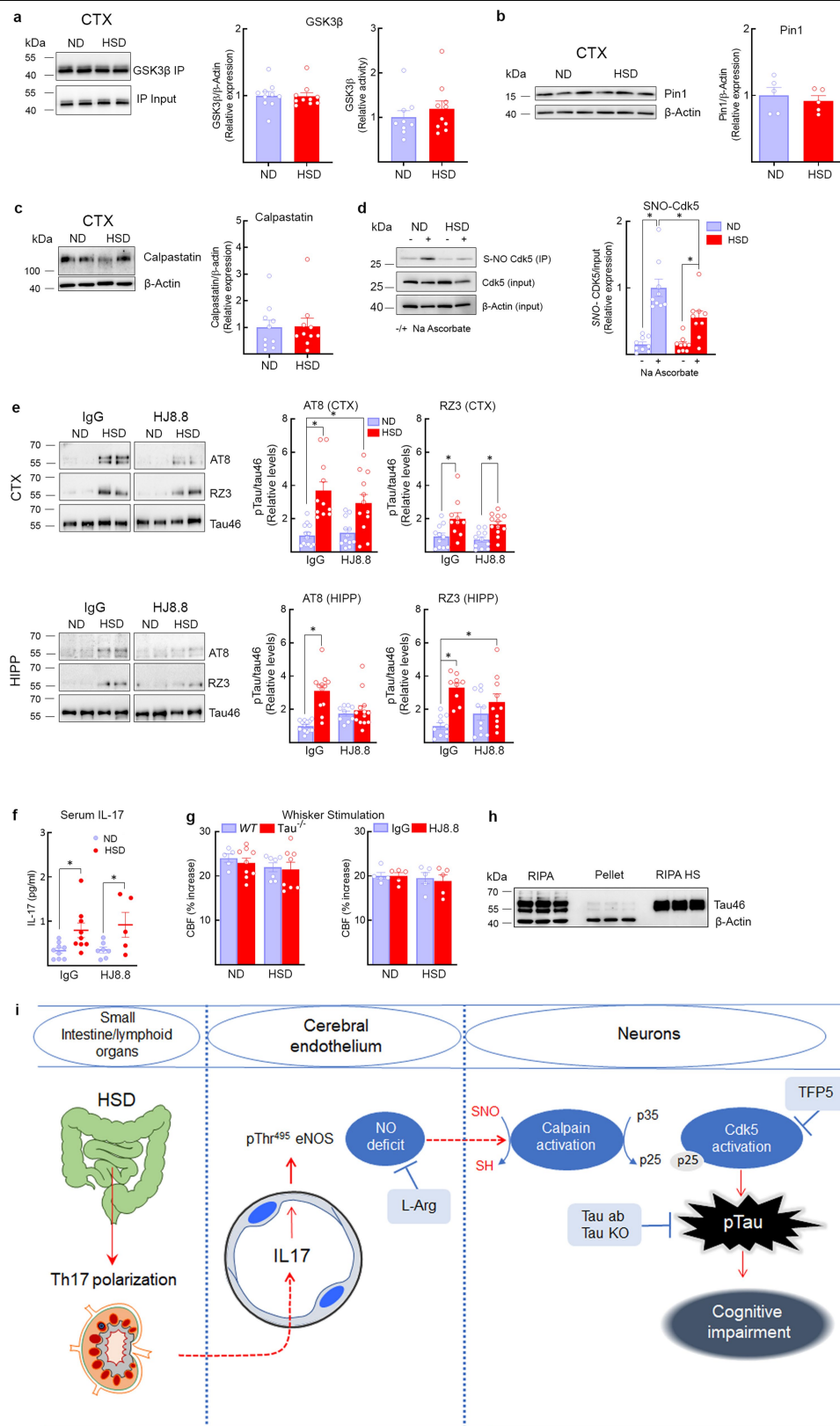
**Extended Data Fig. 3 | A $\beta$  levels in HSD-fed mice and correlation of behavioural deficits with p-tau, as well as p-tau in hypertension, HSD-treated tg2576 mice and hypothermia.** **a**, HSD (8% NaCl) does not alter the distance travelled before finding the escape hole in the Barnes maze (primary distance, ND/HSD  $n=13/13$ , diet:  $*P=0.0462$ , time:  $*P<0.0001$ , two-way repeated measures ANOVA and Bonferroni's test; primary distance day 5: ND/HSD  $n=13/13$ ,  $P=0.0670$  versus ND, two-tailed unpaired  $t$ -test) or the number of errors made (primary errors, ND/HSD  $n=13/13$ , diet:  $P=0.110$ , time:  $*P=0.0004$ , two-way repeated measures ANOVA and Bonferroni's test; primary errors day 5:  $P=0.1226$  versus ND, two-tailed unpaired  $t$ -test). **b**, RZ3 levels in the cortex correlate with cognitive performance on the NOR test. No correlation was found between hippocampal RZ3 levels and cognitive performance on either the Barnes maze or the NOR test. RZ3 cortex: Barnes maze  $r=0.2828$ ,  $*P=0.0133$ ,  $n=76$ ; NOR  $r=-0.2806$ ,  $*P=0.0170$ ,  $n=72$ ; RZ3 hippocampus: Barnes maze  $r=0.1739$ ,  $P=0.1470$ ,  $n=71$ ; NOR  $r=-0.1746$ ,  $P=0.1577$ ,  $n=67$ , Pearson's correlation coefficient). **c**, HSD does not increase soluble or insoluble A $\beta_{38}$ , A $\beta_{40}$  or A $\beta_{42}$  in the neocortex. A $\beta_{38}$ , soluble ND/HSD  $n=11/9$ , insoluble ND/HSD  $n=7/6$ ; A $\beta_{40}$ , soluble ND/HSD  $n=11/14$ , insoluble ND/HSD  $n=7/6$ ; A $\beta_{42}$ , soluble ND/HSD  $n=9/9$ , insoluble ND/HSD  $n=7/6$ . **d**, Delivery of angiotensin II (ANGII; 600 ng kg<sup>-1</sup> min<sup>-1</sup>, subcutaneously (s.c.)) via osmotic minipumps over 6 weeks increases systolic blood pressure (SBP) and induces cognitive deficits

(SBP: Veh/ANGII  $n=10/10$ , treatment:  $*P<0.0001$ , time:  $*P<0.0001$ , repeated measures two-way ANOVA and Bonferroni's test; NOR: 2 weeks Veh/ANGII  $n=12/12$ , 4 weeks Veh/ANGII  $n=10/11$ , 6 weeks Veh/ANGII  $n=7/7$ , treatment:  $*P<0.0021$ , time:  $*P=0.0208$ , two-way ANOVA and Bonferroni's test). **e**, Administration of angiotensin II increases AT8 and RZ3 immunoreactivity in the neocortex but not the hippocampus (cortex, AT8 6 weeks: Veh/ANGII  $n=4/4$ ,  $*P=0.0324$ ; RZ3 6 weeks: Veh/ANGII  $n=5/5$ ,  $*P=0.0262$ ; hippocampus, AT8 6 weeks: Veh/ANGII  $n=5/5$ ,  $P=0.4056$ ; RZ3 6 weeks: Veh/ANGII  $n=5/5$ ,  $P=0.0556$ , two-tailed unpaired  $t$ -test versus vehicle). **f**, HSD increases AT8 and RZ3 levels in both the neocortex and the hippocampus of 6-month-old Tg2576 mice (cortex, AT8:  $*P<0.0001$ ; hippocampus, AT8:  $*P=0.0153$ ; cortex, RZ3:  $*P<0.0001$ ; hippocampus, RZ3:  $*P=0.0239$ ; two-tailed unpaired  $t$ -test for HSD versus ND). **g**, Hypothermia induces massive AT8 phosphorylation (cortex: AT8  $n=4/5$ ,  $*P=0.0159$ ; hippocampus: AT8  $n=4/5$ ,  $*P=0.0159$ ) and increases MC1 (cortex: MC1  $n=4/5$ ,  $*P=0.0317$ ; hippocampus: MC1  $n=4/5$ ,  $*P=0.0159$ ) and RZ3 (cortex: RZ3  $n=4/5$ ,  $*P=0.0201$ ; hippocampus: RZ3  $n=4/5$ ,  $*P=0.0453$ ). Unpaired two-tailed  $t$ -test for hypothermia (HYPO) versus normal temperature (NT). **h**, Unlike HSD (Fig. 1G), hypothermia does not shift tau from soluble to more insoluble fractions. For gel source data see Supplementary Fig. 1. Data are expressed as mean  $\pm$  s.e.m.



**Extended Data Fig. 4 | Effect of L-arginine on p-tau and calpain expression, as well as p-tau in eNOS<sup>-/-</sup> mice, calpain and CDK5 localization, pDARPP-32 with HSD, and IL-17 levels. a**, Administration of L-arginine (10 g l<sup>-1</sup> in drinking water), starting at week 8 of HSD and continuing through week 12, suppresses RZ3 levels in the neocortex but not in the hippocampus (cortex: RZ3, ND/HSD  $n = 10/10$ ,  $*P < 0.0001$ ; hippocampus: RZ3, ND/HSD  $n = 10/10$ ,  $*P = 0.0005$ , two-tailed unpaired  $t$ -test versus normal diet with vehicle). **b**, L-Arginine does not affect the increase in serum IL-17 induced by HSD (Veh, ND/HSD  $n = 9/11$ ,  $*P = 0.0002$  versus ND Veh; L-arg, ND/HSD  $n = 9/8$ ,  $*P < 0.0001$  versus ND L-arg, two-tailed unpaired  $t$ -test). **c**, AT8 and RZ3 levels are elevated in the neocortex and hippocampus of eNOS<sup>-/-</sup> mice on ND (AT8: cortex, ND/HSD  $n = 5/4$ ,  $*P = 0.0029$ ; hippocampus, ND/HSD  $n = 5/4$ ,  $*P = 0.0078$ ; RZ3: cortex, ND/HSD  $n = 5/4$ ,  $*P = 0.0003$ ; hippocampus, ND/HSD  $n = 5/4$ ,  $*P = 0.0128$ , two-tailed unpaired  $t$ -test versus wild-type mice). **d**, HSD does not increase tau

phosphorylation in eNOS<sup>-/-</sup> mice (RZ3: hippocampus, ND/HSD  $n = 7/8$ ,  $*P = 0.0224$  versus ND, two-tailed unpaired  $t$ -test). **e**, Calpain 2 immunoreactivity is present in neuronal cell bodies of the somatosensory and piriform cortex (scale bars, 500  $\mu$ m (left); 100  $\mu$ m (right)). Representative images from  $n = 3$  mice. **f**, Colocalization of Calpain 2 and CDK5 in neuronal cell bodies of the piriform cortex (scale bars, 50  $\mu$ m (main images); 10  $\mu$ m (inset)). Representative images from  $n = 3$  mice. **g**, HSD has no effect on the phosphorylation of the CDK5 substrate DARPP-32 in neocortex; ND/HSD  $n = 10/10$ . **h**, Administration of the CDK5 peptide inhibitor TFP5 has no effect on the increase in serum IL-17 levels induced by HSD (scrambled: ND/HSD  $n = 5/4$ ,  $*P = 0.0002$  versus ND scrambled; TFP5: ND/HSD  $n = 7/8$ ,  $*P < 0.359$  versus ND TFP5; two-tailed unpaired  $t$ -test). **i**, L-Arginine does not alter the levels of calpain 1 and 2 in the neocortex or hippocampus. ND/HSD  $n = 3/5$ . For gel source data see Supplementary Fig. 1. Data are expressed as mean  $\pm$  s.e.m.



**Extended Data Fig. 5** | See next page for caption.

**Extended Data Fig. 5 | GSK3 $\beta$ , PIN-1, calpastatin and CDK5 nitrosylation in HSD-fed mice, as well as neurovascular coupling, effect of HJ8.8 on p-tau, serum IL-17, and summary.** **a**, HSD has no effect on the expression or activity of GSK3 $\beta$  in the neocortex. ND/HSD  $n = 10/10$ . **b**, HSD does not alter the expression of the prolyl *cis/trans* isomerase PIN-1, a regulator of tau dephosphorylation. ND/HSD  $n = 5/5$ . **c**, The expression of calpastatin, an endogenous inhibitor of calpain activity, is not reduced by HSD. ND/HSD  $n = 10/10$ . **d**, Nitrosylation of CDK5 is reduced in the neocortex of HSD-fed mice (ND/HSD  $n = 9/9$ , diet:  $*P = 0.0143$ ; ascorbate:  $*P < 0.0001$ ; two-way ANOVA and Tukey's test). **e**, HJ8.8 reduces AT8 in the hippocampus (IgG: ND/HSD  $n = 13/12$ ; HJ8.8: ND/HSD  $n = 9/13$ ;  $*P < 0.0001$ , Kruskal–Wallis test and Dunn's test). RZ3 levels are not altered by HJ8.8. **f**, Administration of HJ8.8 does alter the increase in serum IL-17 levels induced by HSD (IgG: ND/HSD  $n = 9/9$ ,  $*P = 0.0192$  versus ND IgG; HJ8.8: ND/HSD  $n = 7/5$ ,  $*P = 0.0421$  versus ND HJ8.8, two-tailed unpaired *t*-test). **g**, The increase in somatosensory cortex CBF induced by neural activity evoked by mechanical stimulation of the whiskers is not reduced by HSD in wild-type, tau<sup>-/-</sup> or HJ8.8-

treated mice (wild-type ND/HSD  $n = 5/7$ , tau<sup>-/-</sup> ND/HSD  $n = 9/8$ ; IgG ND/HSD  $n = 5/5$ , HJ8.8 ND/HSD  $n = 5/5$ ). **h**, Western blotting showing enrichment of tau in boiled RIPA neocortical samples (heat-stable fraction, HS). Note that  $\beta$ -actin is lost during the boiling process. Representative images from  $n = 3$  experiments. **i**, Cartoon depicting the mechanisms by which HSD leads to tau phosphorylation and cognitive impairment. HSD elicits a T<sub>H</sub>17 response in the small intestine, which leads to an increase in circulating IL-17. IL-17, in turn, suppresses endothelial NO production by inducing inhibitory phosphorylation of eNOS at Thr495. The NO deficit results in reduced nitrosylation of calpain in neurons, and increases in calpain activity, p35 to p25 cleavage, activation of CDK5, and tau phosphorylation, which is ultimately responsible for cognitive dysfunction. In support of this chain of events, rescuing the endothelial NO deficit with L-arginine, lack of tau in tau-null mice, treatment with the CDK5 peptide inhibitor TFP5 or treatment with antibodies directed against tau (Tau ab) prevent the cognitive dysfunction. For gel source data see Supplementary Fig. 1. Data are expressed as mean  $\pm$  s.e.m.



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated
- ☐ ☒ Clearly defined error bars  
*State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on [statistics for biologists](#) may be useful.

### Software and code

Policy information about [availability of computer code](#)

#### Data collection

Chart 5 Pro (v5.5.6) was used for collection of CBF and MAP data. Any Maze (v5.3) was used for collection of behavioral data. IPLab-2.8.0 was used for acquiring phospho-Tau immunostaining images. Microsoft Excel (for Office365) was used for mouse randomization.

#### Data analysis

Chart 5 Pro (v5.5.6) was used for analysis of CBF data. Any Maze (v5.3) was used for analysis of behavioral data. Biorad Image Lab (v6.0) was used for analysis of immunoblots. Image J (v1.52p) was used for analysis of ASL MRI data. Graph Pad (v8.0) software was used for statistical analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated or analysed during this study are included in this published article (and its supplementary information files).

## Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined according to power analysis based on previous published works published by our lab on CBF regulation and behavior.
Data exclusions	No data were excluded.
Replication	At least three independent experiments (with a number of subjects ranging from 3 to 5 each) were performed for all the findings in the manuscript. All attempts at replication were successful and are included in the figures.
Randomization	Mouse randomization was based on the random number generator function (RANDBETWEEN) in Microsoft Excel software.
Blinding	Analysis was performed in a blinded fashion.

## Reporting for specific materials, systems and methods

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input type="checkbox"/>	<input checked="" type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used

pTau Ser199Ser202, Rabbit polyclonal (ThermoFisher #44-768G); pTau Ser202Thr205 (AT8), Mouse monoclonal (ThermoFisher #MN1020); pTau Ser396 (PHF13), Mouse monoclonal (Cell Signaling #9632); pThr231 (RZ3) Mouse monoclonal (Gift from Prof. Peter Davies); MC-1 Mouse monoclonal (Gift from Prof. Peter Davies); K280, Rabbit polyclonal (Anaspec #AS-56077); pSer396 (PHF1), Mouse monoclonal (Gift from Prof. Peter Davies); Tau46, Mouse monoclonal (Cell Signaling #4019); p35/p25, Rabbit monoclonal, Clone C64B10 (Cell Signaling #2680); Cdk5, Rabbit polyclonal (Cell Signaling #2506); GSK3 $\beta$ , Rabbit polyclonal, Cell Signaling (#9315); Calpain 1, Rabbit polyclonal (Cell Signaling #2556); Calpain 2, Rabbit polyclonal (Cell Signaling #2539); Calpain 2 (IHC), Mouse monoclonal, E10 (Santa Cruz #373966); Calpastatin, Rabbit polyclonal (Cell Signaling #4146); pThr75 DARPP-32, Rabbit polyclonal (Cell Signaling #2301); DARPP-32 Rabbit polyclonal (Cell Signaling #2302); Pin1, Rabbit polyclonal (Cell Signaling #3722);  $\beta$ -Actin, Mouse monoclonal (Sigma #A5441); NeuN, Mouse monoclonal (Millipore #MAB377); Iba1, Rabbit polyclonal (Wako #WEG2172); GFAP, Mouse monoclonal, clone G-A-5 (Sigma #63893); N/CD13, Goat polyclonal (R&D Systems #AF2335).

## Validation

AT8 and pSer199Ser202 Tau antibody were validated by treating the brain samples with Lambda phosphatase. All phospho-Tau antibodies and the K280 antibody were also validated in mice with deletion of Tau. Cdk5 and p35p25 antibodies have been validated in immuno-precipitation experiments with IgG control antibody. All the other antibodies were extensively validated by us and/or in the literature.

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

## Laboratory animals

Studies were conducted, according to the ARRIVE guidelines (<https://www.nc3rs.org.uk/arrive-guidelines>), in the following lines of male mice (age: 8weeks): C57BL/6 (JAX), B6.129X1-Mapt<sup>tm1Hnd</sup> (Tau <sup>-/-</sup>, JAX, Stock #007251), Tg(Camk2a-tTA)1Mmay Fgf14Tg(tetO-MAPT\*P301L)4510Kha/J (rTg4510, JAX, Stock#024854), B6.129P2-Nos3tm1Unc/J (eNOS<sup>-/-</sup>, JAX, Stock #002684), B6.129S4-Nos1tm1Plh/J (nNOS<sup>-/-</sup>, JAX, Stock #002986), and 129S6.Cg-Tg(APP<sup>SWE</sup>)2576Kha N20+? (Tg2576 or APP<sup>SWE</sup>, Taconic, Stock #279).

## Wild animals

The study did not involve wild animals

## Field-collected samples

The study did not involve field-collected samples.

## Magnetic resonance imaging

### Experimental design

## Design type

Indicate task or resting state; event-related or block design.

## Design specifications

Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.

## Behavioral performance measures

State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

### Acquisition

## Imaging type(s)

Arterial Spin Labeling MRI

## Field strength

7.0 Tesla 70/30 Bruker Biospec small-animal MRI system with 450 mT/m gradient amplitude and a 4500 T · m<sup>-1</sup> · s<sup>-1</sup> slew rate.

## Sequence &amp; imaging parameters

ASL-MRI was based on a flow-sensitive alternating inversion recovery rapid acquisition with relaxation enhancement (FAIR-RARE) pulse sequence labeling the inflowing blood by global inversion of the equilibrium magnetization.

## Area of acquisition

The resting cerebral blood flow was measured in the cortex and in the hippocampus.

## Diffusion MRI

☐ Used

☒ Not used

### Preprocessing

## Preprocessing software

Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).

## Normalization

If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.

## Normalization template

Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.

## Noise and artifact removal

Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).

## Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

### Statistical modeling & inference

## Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

## Effect(s) tested

We tested the effect of high salt diet on the resting CBF. Specifically, we tested whether anti-tau antibodies restored the resting CBF reduction mediated by high salt diet.

## Specify type of analysis:

☐

Whole brain

☒

ROI-based

☐

Both

Anatomical location(s)

Describe how anatomical locations were determined (e.g. specify whether automated labeling algorithms or probabilistic atlases were used).

Statistic type for inference  
(See [Eklund et al. 2016](#))

The images were analyzed by Image J and the average CBF value is reported as mL per 100g of tissue per minute.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Functional and/or effective connectivity
<input checked="" type="checkbox"/>	<input type="checkbox"/> Graph analysis
<input checked="" type="checkbox"/>	<input type="checkbox"/> Multivariate modeling or predictive analysis

# Cyclic GMP–AMP signalling protects bacteria against viral infection

<https://doi.org/10.1038/s41586-019-1605-5>

Received: 25 June 2019

Accepted: 11 September 2019

Published online: 18 September 2019

Daniel Cohen<sup>1,3</sup>, Sarah Melamed<sup>1,3</sup>, Adi Millman<sup>1</sup>, Gabriela Shulman<sup>1</sup>,  
Yaara Oppenheimer-Shaanan<sup>1</sup>, Assaf Kacen<sup>2</sup>, Shany Doron<sup>1</sup>, Gil Amitai<sup>1\*</sup> & Rotem Sorek<sup>1\*</sup>

The cyclic GMP–AMP synthase (cGAS)–STING pathway is a central component of the cell-autonomous innate immune system in animals<sup>1,2</sup>. The cGAS protein is a sensor of cytosolic viral DNA and, upon sensing DNA, it produces a cyclic GMP–AMP (cGAMP) signalling molecule that binds to the STING protein and activates the immune response<sup>3–5</sup>. The production of cGAMP has also been detected in bacteria<sup>6</sup>, and has been shown, in *Vibrio cholerae*, to activate a phospholipase that degrades the inner bacterial membrane<sup>7</sup>. However, the biological role of cGAMP signalling in bacteria remains unknown. Here we show that cGAMP signalling is part of an antiphage defence system that is common in bacteria. This system is composed of a four-gene operon that encodes the bacterial cGAS and the associated phospholipase, as well as two enzymes with the eukaryotic-like domains E1, E2 and JAB. We show that this operon confers resistance against a wide variety of phages. Phage infection triggers the production of cGAMP, which—in turn—activates the phospholipase, leading to a loss of membrane integrity and to cell death before completion of phage reproduction. Diverged versions of this system appear in more than 10% of prokaryotic genomes, and we show that variants with effectors other than phospholipase also protect against phage infection. Our results suggest that the eukaryotic cGAS–STING antiviral pathway has ancient evolutionary roots that stem from microbial defences against phages.

Bacterial antiphage immune systems, such as CRISPR–Cas and restriction modification systems, tend to concentrate in ‘defence islands’ in bacterial genomes<sup>8</sup>; this property has facilitated the discovery of defence systems on the basis of their colocalization with known ones<sup>9–11</sup>. We noticed that homologues of the gene that encodes cGAS in *V. cholerae* (*dncV*; ‘dinucleotide cyclase in *Vibrio*’) frequently tend to appear near defence genes. Out of 637 homologues of this protein that we identified through a homology search in 38,167 microbial genomes, we found that 417 (65.5%) are located in the vicinity of known defence systems (Fig. 1a). It has previously been shown that such a high propensity of colocalization with defence genes is a strong predictor that the gene under inspection has a role in phage resistance<sup>11</sup>. These results therefore suggest that the gene that encodes the bacterial cGAS participates in defence against phages.

The bacterial cGAS and the effector phospholipase ‘cGAMP-activated phospholipase in *Vibrio*’ (CapV)<sup>7</sup> are encoded by adjacent genes in the *V. cholerae* genome and are probably expressed in a single operon<sup>7</sup>. The same operon also contains two additional genes, the presence of which next to the *capV-dncV* gene pair is conserved in the majority of cases (96%, 613 out of 637 homologues), suggesting that the putative functional defence system that involves bacterial cGAS comprises these four genes (Fig. 1a). As has previously been noted<sup>7,12</sup>, the two additional genes encode proteins with domains that are known to be associated with the eukaryotic ubiquitin system: the E1 and E2 domains that are typical of ubiquitin transfer and ligation enzymes, and a JAB domain,

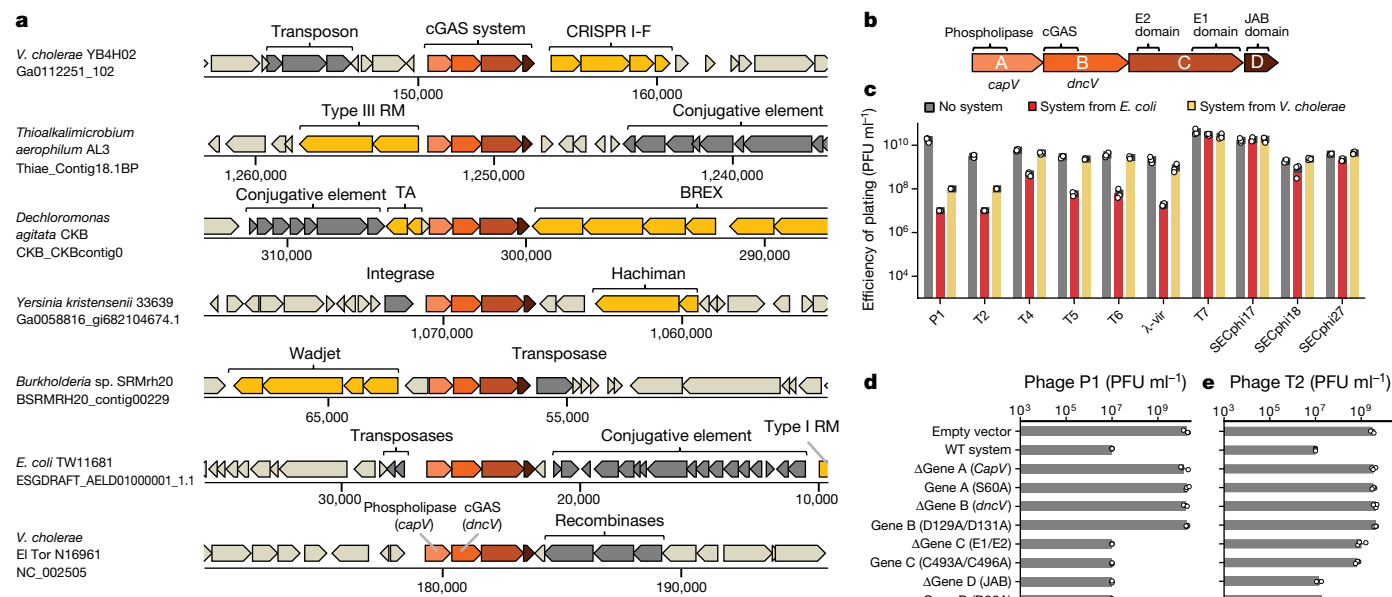
which is similar to de-ubiquitinase enzymes that remove ubiquitin from target proteins (Fig. 1b).

To test whether the four-gene operon that contains cGAS is an antiphage defence system, we cloned the operon of *V. cholerae* serovar O1 biovar El Tor (hereafter, *V. cholerae* El Tor) or the homologous operon from *Escherichia coli* strain TW11681 into the laboratory strain *E. coli* MG1655, which naturally lacks this system. In both cases, the putative four-gene system was cloned together with its upstream and downstream intergenic regions to preserve promoters, terminators and other regulatory sequences. We then challenged *E. coli* MG1655 bacteria containing these four-gene systems with an array of phages that spans the three major families of tailed, double-stranded DNA phages (T2, T4, T6 and P1 from the *Myoviridae*;  $\lambda$ -vir, T5, SECphi18 and SECphi27 from the *Siphoviridae*; and T7 from the *Podoviridae*), as well as the single-stranded DNA phage SECphi17, of the *Microviridae* family.

Both the *V. cholerae*-derived and the *E. coli*-derived four-gene operons conferred defence against multiple phages. The system from *E. coli* provided 10–1,800-fold protection against 6 of the 10 phages that we tested, and the system from *V. cholerae* protected against 2 of the phages (Fig. 1c, Extended Data Fig. 1). None of the systems protected against the transformation of a multi-copy plasmid (Extended Data Fig. 2). The stronger protective effect that was observed for the *E. coli*-derived system, as compared to that from *V. cholerae*, possibly stems from the higher compatibility of the former with the *E. coli* host and the coliphages

<sup>1</sup>Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel. <sup>2</sup>Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. <sup>3</sup>These authors contributed equally: Daniel Cohen, Sarah Melamed. \*e-mail: gil.amitai@weizmann.ac.il; rotem.sorek@weizmann.ac.il





**Fig. 1 | Systems containing bacterial cGAS protect against phage infection.**

**a**, Genes that encode bacterial cGAS are part of a conserved four-gene operon that is genomically associated with antiphage defence systems. Representative instances of the four-gene putative defence systems that contain cGAS, and their genomic environments, are presented. Genes that are known to be involved in defence are shown in yellow. Genes of mobile genetic elements are in dark grey. RM, restriction modification; TA, toxin–antitoxin. BREX, Hachiman and Wadjet are recently described defence systems<sup>9,11</sup>. **b**, Domain organization of the genes in the operon that contains cGAS. **c**, The four-gene operon from *V. cholerae* El Tor or *E. coli* TW11681 was cloned into *E. coli* MG1655 (Methods). The efficiency of plating is shown for ten phages infecting the control *E. coli* MG1655 strain (no system), the strain with the four-gene operon cloned from *E. coli* TW11681

(system from *E. coli*) and the strain with the four-gene operon cloned from *V. cholerae* El Tor (system from *V. cholerae*). Data represent plaque-forming units (PFU) per millilitre; bar graph represents average of three independent replicates, with individual data points overlaid. **d**, **e**, Efficiency of plating of phages infecting strains with the wild-type (WT) four-gene operon cloned from *E. coli* TW11681, deletion strains in that operon and strains with point mutations. Data represent plaque-forming units per millilitre; bar graph represents average of three independent replicates, with individual data points overlaid. Empty vector represents a control *E. coli* MG1655 strain that lacks the system and contains an empty vector instead. **d**, Infection with phage P1. **e**, Infection with phage T2.

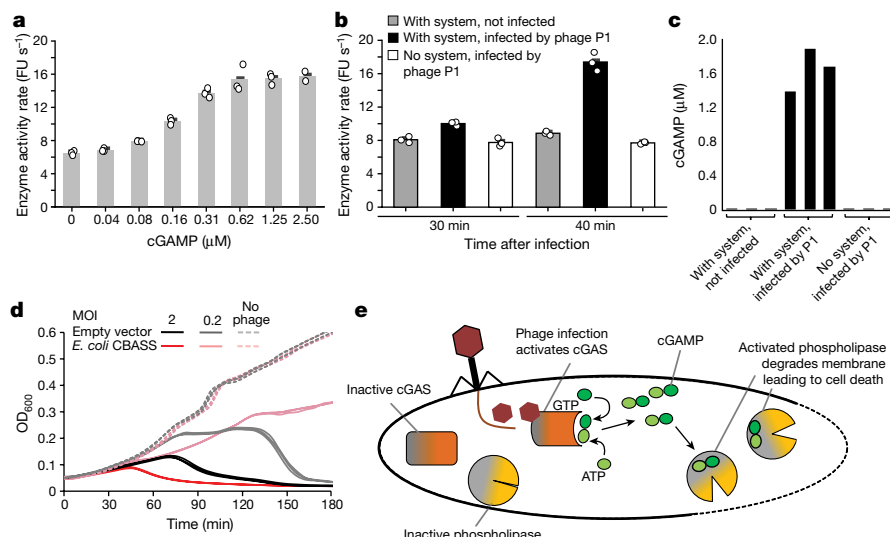
that we used. We therefore proceeded with the *E. coli*-derived system for further experiments.

To determine whether the ability to produce and sense cGAMP affects defence, we experimented with mutated forms of the system. Deletion of either the cGAS-encoding or the phospholipase-encoding genes resulted in a complete loss of protection against phage infection (Fig. 1d, e, Extended Data Fig. 3a–d). Moreover, point mutations in two essential aspartate residues in the cGAMP-producing active site of cGAS (D129A/D131A)<sup>7,13</sup> abolished defence, which suggests that cGAMP production is absolutely necessary for the phage-defensive properties of the system (Fig. 1d, e, Extended Data Fig. 3a–d). A S60A point mutation that inactivates the catalytic site of the CapV phospholipase also rendered the system completely inactive against all the phages that we tested, demonstrating that the cGAMP-controlled phospholipase activity of CapV is essential for defence (Fig. 1d, e, Extended Data Fig. 3a–d). These results indicate that, as in eukaryotes, the cGAS–cGAMP signalling pathway in bacteria participates in antiviral defence.

We next examined mutations in the two additional genes in the four-gene operon. Defence against the myophage P1 was not affected when the gene encoding the E1 and E2 domains or the gene encoding the JAB domain were deleted, which suggests that these two genes are not essential for protection against P1 (Fig. 1d). Accordingly, a construct that contained only the *capV*–*dncV* gene pair (that is, expressed only the phospholipase–cGAS pair) showed complete defence against P1 (Extended Data Fig. 4), demonstrating that the phospholipase–cGAS pair can work as a standalone defence system against some phages. However, deletion of the gene that encodes the E1 and E2 domains abolished defence against all of the other phages that were tested (T2, T4, T5, T6 and λ-vir). The presence of an intact gene that encodes the JAB-domain protein was necessary for protection against some phages

(T4, T5 and T6), but not against others (P1, T2 and λ-vir). Point mutations that are predicted to inactivate the active site of the E1 domain (C493A/C496A) or to inactivate the active site of the peptidase in the JAB domain (D38A) recapitulated the results of deletions of the entire genes, suggesting that the predicted enzymatic activities of the E1–E2 and JAB proteins are necessary for their roles in phage defence (Fig. 1d, e, Extended Data Fig. 3a–d). Our mutational analyses suggest that the four-gene operon forms a bacterial system that relies on cGAMP signalling and phospholipase activity to defend against a broad range of phages, and that the activities of the two additional genes are necessary for defence against some—but not all—phages. We denote this system the cyclic-oligonucleotide-based antiphage signalling system (CBASS).

In animals, the sensing of viral infection is the trigger that activates the production of cGAMP<sup>3</sup>. We therefore sought to examine whether phage infection triggers the production of cGAMP in the bacterial system as well. To this end, we took advantage of the fact that the CapV phospholipase is activated by cGAMP, and used this protein as a reporter for the presence of cGAMP in infected cells. We first expressed and purified the CapV phospholipase from the *E. coli* CBASS system, and measured its activity using an in vitro phospholipase activity assay. As expected, chemically synthesized 3′/3′-cGAMP induced CapV phospholipase activity in a concentration-dependent manner (Fig. 2a). Next, we exposed the purified CapV to cell lysates that were derived from cells containing the Δ*capV* CBASS, and which had been filtered to include small molecules only (Methods). The CapV phospholipase became markedly activated when exposed to cell lysates that were collected 40 min after infection with phage P1, which suggests that phage infection triggers cGAMP accumulation in cells that contain the CBASS (Fig. 2b). These results were corroborated by targeted mass spectrometry analysis, which detected cGAMP concentrations of 1.4–1.9 μM in lysates that were



**Fig. 2 | Phage infection triggers cGAMP accumulation and cell death.**

**a**, Purified CapV protein was incubated in vitro with synthetically produced 3′3′-cGAMP in the presence of resorufin butyrate, a phospholipase substrate that emits fluorescence when hydrolysed. The x axis shows the concentration of cGAMP added (μM); the y axis shows the enzyme activity rate, measured by the accumulation rate of fluorescence units (FU) per second (Methods). Bar graph represents average of three technical replicates, with individual data points overlaid (except for the 2.5 μM concentration, for which the data represent an average of two replicates). **b**, Purified CapV protein was incubated in vitro with filtered cell lysates derived from bacterial cultures infected by phage P1 at a multiplicity of infection (MOI) of 2. Lysates were extracted from cells containing the CBASS with the *capV* gene deleted (‘with system’) or from control cells that lack the CBASS and contain an empty vector instead (‘no system’). Lysates were collected 30 min or 40 min after initial infection. Bar graph

represents average of three technical replicates, with individual data points overlaid. **c**, Targeted mass spectrometry analysis was performed on the filtered cell lysates collected 40 min after initial infection by phage P1 at an MOI of 2. Lysates from uninfected samples were taken as control. The y axis represents the concentration of 3′3′-cGAMP in the cell lysate measured by mass spectrometry, as calculated on the basis of a calibration curve with synthetically produced 3′3′-cGAMP (Methods). Three independent replicates for each condition are shown, and each bar represents an individual replicate. **d**, Growth curves in liquid culture for bacteria that contain the CBASS and bacteria that lack the CBASS (empty vector), infected by phage P1 at 37 °C. Bacteria were infected at time 0 at an MOI of 0.2 or 2. OD<sub>600</sub>, optical density at a wavelength of 600 nm. Three independent replicates are shown for each MOI, and each curve shows an individual replicate. **e**, Model for CBASS-based antiphage activity in bacteria.

collected 40 min after infection (Fig. 2c). CapV was not activated when exposed to lysates that were collected from uninfected cells, and only slightly activated by lysates that were collected from cells 30 min after infection. This implies that the phage component or cell state sensed by the CBASS appears in the cell relatively late in the phage-infection cycle (Fig. 2b).

It has previously been shown that, when stimulated by cGAMP, the CapV phospholipase degrades the membrane of *V. cholerae*, which leads to a loss of membrane integrity and to the arrest of cell growth or to death<sup>7</sup>. This, combined with our finding that the production of cGAMP is triggered by phage infection, led us to hypothesize that this defence system executes its defence via abortive infection. Abortive-infection defence systems exert their activity by causing the infected bacterial cell to ‘commit suicide’ before the phage replication cycle can be completed. This strategy eliminates infected cells from the bacterial population and protects the culture from a viral epidemic<sup>14</sup>. A phenotype such as this predicts that, with a high multiplicity of infection (MOI) (in which nearly all bacteria are infected in the first cycle), massive cell death will be observed in the culture, even for cells that contain the defence system.

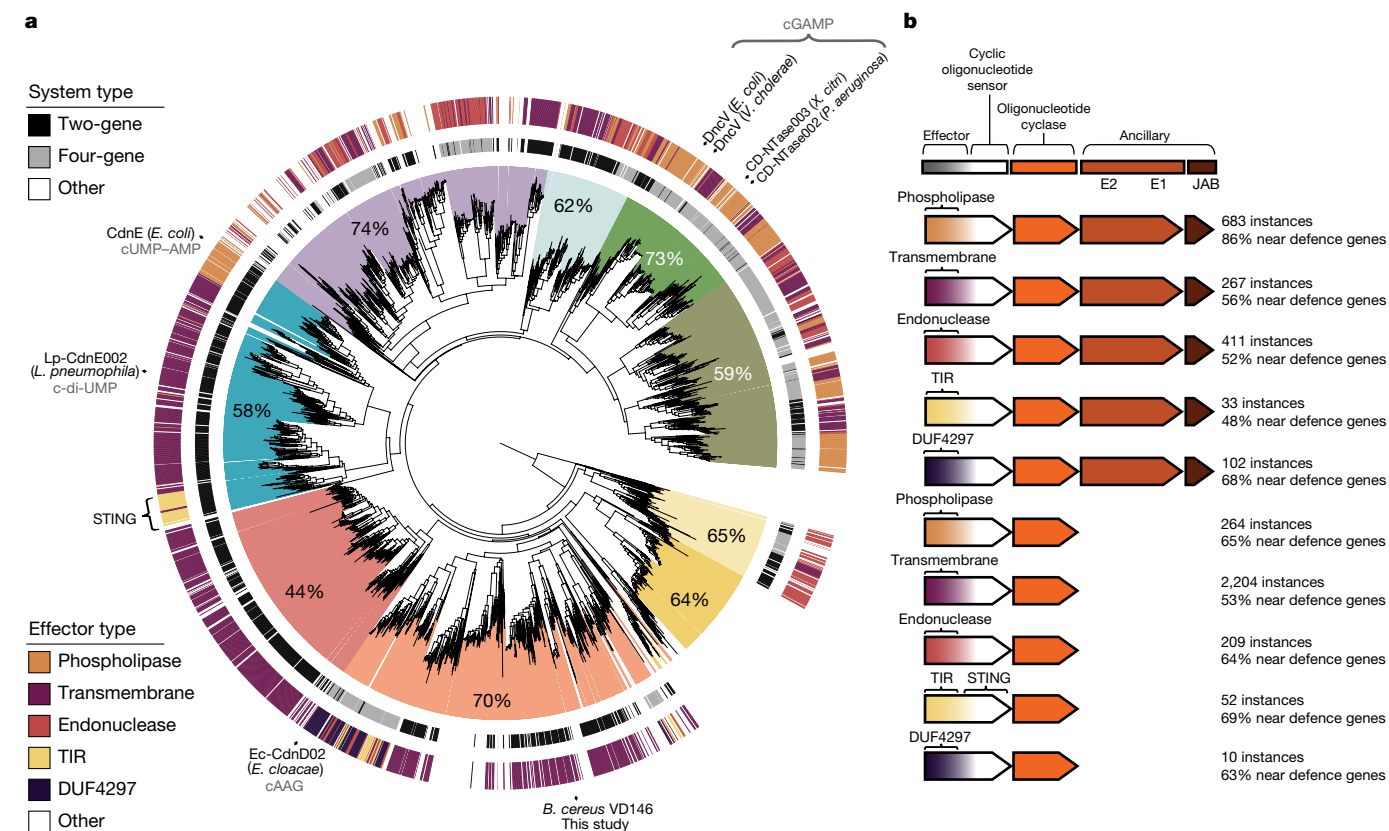
To test this hypothesis, we infected bacteria growing in liquid cultures with phage P1 at varying MOIs and examined the culture dynamics. At an MOI of 0.2 (in which only around 20% of bacteria are initially infected by the phage) the culture of the wild-type cells collapsed owing to phage propagation, whereas the culture of the cells containing the CBASS was viable (Fig. 2d). Conversely, at an MOI of 2 (in which almost all bacteria are infected by the phage), the culture of wild-type cells and the culture of cells containing the CBASS both collapsed, which indicates cell lysis of the infected, CBASS-containing cells. Moreover, whereas at an MOI of 2 the wild-type culture started collapsing 70 min after initial infection (consistent with a time to lysis of 60–70 min after infection, as has

previously been reported for phage P1<sup>15</sup>), the culture that contained the CBASS started collapsing as early as 45 min after infection—a period that is insufficient for the completion of the phage P1 replication cycle (Fig. 2d). These results were reproduced for other phages; when phages were applied at a high MOI, the cells that contained the CBASS showed lysis before the necessary time had elapsed for the completion of the phage replication cycle (Extended Data Fig. 5a). In the case of phage P1, the same phenotype was maintained for constructs that included only *capV-dncV* gene pair (that is, without the genes encoding the E1–E2 domains and JAB domain proteins), indicating that the latter two genes are not necessary for defence by abortive infection against phage P1 (Extended Data Fig. 5b).

These results were further corroborated by the staining of infected cells with propidium iodide, a fluorescent DNA-binding agent that is not membrane-permeable and cannot normally enter cells. Fluorescence-activated cell sorting (FACS) analysis showed that at 40 min after infection by phage P1 a substantial population of the cells that contain the CBASS became stained by propidium iodide, which indicates a loss of membrane integrity (Extended Data Fig. 6). This was associated with the loss of normal cell shape, as was observed via microscopy (Extended Data Fig. 7).

The above results suggest a model in which the cGAS–phospholipase defence system somehow senses a phage infection, and this sensing then triggers the production of cGAMP by the cGAS protein. The cGAMP molecule in turn activates the phospholipase, which degrades the bacterial membrane and causes cell death (Fig. 2e). This cell death before the completion of the phage replication cycle aborts the infection and prevents further propagation of the phage.

It has recently been shown that the bacterial cGAS gene (*dncV*) belongs to a large family of oligonucleotide cyclases, members of which are



**Fig. 3 | Widespread occurrence of CBASSs in prokaryotic genomes.**

**a**, Phylogenetic tree of proteins with predicted oligonucleotide cyclase domains that were identified by searching a set of 38,167 microbial genomes (Methods). Clades are coloured following a previous colour coding<sup>16</sup>. The percentage of genes that are located near known defence systems is indicated for each clade. The outermost ring indicates the type of effector gene that is associated with the oligonucleotide cyclase gene; bottom left, colour code for the effector types. The inner ring indicates the type of system: the two-gene system (comprising an oligonucleotide cyclase and an effector) or four-gene system (that also includes

genes encoding the E1–E2 domains and JAB domain proteins). Top left, colour code for system type. Selected oligonucleotide cyclase genes that were biochemically characterized<sup>16</sup> are indicated; the cyclic oligonucleotide that each produces appears in grey font. cAAG, cyclic trinucleotide AMP–AMP–GMP. *E. cloacae*, *Enterobacter cloacae*; *L. pneumophila*, *Legionella pneumophila*; *P. aeruginosa*, *Pseudomonas aeruginosa*; *X. citri*, *Xanthomonas citri*. **b**, Common configurations of four-gene and two-gene CBASSs. The identified number of instances of each such system, and the percentage occurrence next to known defence systems, are indicated for each configuration.

found in about 10% of all sequenced bacterial genomes<sup>16</sup>. Other genes in this family synthesize a diverse set of cyclic oligonucleotide molecules, including cyclic UMP–AMP, cyclic di-UMP and even the cyclic trinucleotide AMP–AMP–GMP<sup>16</sup>. This family of genes has been divided into several clades on the basis of sequence similarity between its members<sup>16</sup>. We found that all major clades of this gene family have a high propensity to be genomically associated with other known defence genes in defence islands (Fig. 3a). Between 44% and 74% of the genes in each clade are located near known defence systems, which suggests that this entire family of oligonucleotide cyclases participates in phage defence (Fig. 3a).

When examining the genomic environment of the 6,232 genes that we identified as belonging to this family (Supplementary Table 1, Methods), we found that in 26% of the cases (1,612 instances) they appeared as part of a four-gene operon that resembled the CBASS of *V. cholerae* (including the genes encoding the E1–E2 domains and the JAB-domain proteins). In 683 (42%) of these operons the predicted effector gene had a phospholipase domain, but in the remaining cases the phospholipase domain in the effector gene was replaced by another domain (Fig. 3b). Common alternative effector domains included endonucleases (usually of the HNH type); a domain comprising predicted transmembrane helices; a domain of unknown function (DUF4297) that has previously been shown to participate in the Lamassu phage-resistance system<sup>11</sup>; and a Toll interleukin receptor (TIR) domain that was also previously shown to participate in antiphage defence<sup>11</sup> (Fig. 3b). These observations imply

that in these variants of the CBASS, the alternative domains replace the phospholipase in exerting the cell-suicide effector activity. For example, the endonuclease may degrade cellular DNA and the transmembrane-domain effector may oligomerize to form membrane pores, as in the case of the RexA–RexB abortive infection system<sup>17</sup>.

In addition, and consistent with previous reports<sup>18</sup>, we found 2,745 cases in which the genes that belong to the family of oligonucleotide cyclases appeared in the context of a two-gene operon that lacked the genes encoding the E1–E2 domains and JAB-domain proteins. The second gene in the operon also frequently included phospholipase, endonuclease or transmembrane helix domains, which suggests that these operons represent a minimal CBASS that comprises only two genes (Fig. 3b). The most common predicted effector domain in these operons was a domain that included either two or four transmembrane helices; effector domains of this type are present in over 2,000 instances of the two-gene operons that we identified.

To test whether predicted two-gene CBASSs have a role in phage resistance, we examined a two-gene system from *Bacillus cereus* VD146 that contains a predicted effector gene with four transmembrane helices (Extended Data Fig. 8a). We engineered this system into the laboratory strain *Bacillus subtilis* BEST7003 and challenged the engineered strain with an array of 11 *Bacillus* phages, as previously described<sup>11</sup> (Methods). The system conferred strong defence against one of these phages (the myophage SBSphC), verifying its ability to defend against phages

(Extended Data Fig. 8b). Deletion of the cGAS-like gene or of the effector gene rendered the system inactive (Extended Data Fig. 8b). Finally, infection assays of bacteria that contained the two-gene system in liquid culture showed culture collapse at a high MOI, consistent with an abortive infection system (Extended Data Fig. 8c).

We observed 52 instances of a two-gene CBASS in which the effector gene contained an N-terminal TIR domain, and also a C-terminal STING domain (Fig. 3b, Extended Data Fig. 9). Such a domain arrangement of STING fused to a TIR domain is also found in primitive eukaryotes, including the oyster *Crassostrea gigas* and the annelid worm *Capitella teleta*<sup>1</sup>. It is therefore possible that the CBASS with the TIR-STING effector gene represents the ancient evolutionary origin of the eukaryotic cGAS-STING system.

This study reveals the biological role of a large family of defence systems that is widespread in microbial genomes, but much remains unknown. The exact phage component sensed by the system is yet to be identified; it is unlikely that this component is cytoplasmic double-stranded DNA (as is the case in the animal cGAS-STING system), because bacteria do not have a nucleus and their cytoplasm therefore always contains double-stranded DNA. The role of the genes encoding the E1-E2 domains and JAB-domain proteins also remains unknown. These genes could be involved in the sensing of some phages, or in the mitigation of anti-cGAS activities that phages are likely to encode.

Accumulating evidence suggests that important components of the eukaryotic innate immune system have counterparts in bacterial immune systems. Argonaute—a central protein in the antiviral RNA interference machinery of plants, insects and animals<sup>19</sup>—has also been reported to have immunity roles in bacteria and archaea<sup>20,21</sup>. TIR domains, which are essential components of the pathogen-recognizing Toll-like receptors<sup>22</sup>, are abundant in bacteria and have recently been shown to have a primary role in antiphage defence<sup>11</sup>. Foreign RNA is sensed in eukaryotic cells by the oligo-adenylate synthase (OAS) protein, leading to the activation of a non-specific RNase<sup>23</sup>; this process has recently been shown to have parallels in type III CRISPR-Cas immunity upon the sensing of phage RNA<sup>24</sup>. All of these processes, in addition to our finding that cGAS signalling in prokaryotes has an antiviral role similar to that in eukaryotes, are unlikely to have been the result of parallel evolution. Instead, these observations cumulatively point to a scenario in which these defence systems first evolved in prokaryotes as means of defence against phages, and the ancient eukaryote (which was probably formed by fusion of a bacterium and an archaeon<sup>25</sup>) inherited a primordial version of these systems from the prokaryotes that formed it. Under this hypothesis, these systems became the basis for the primitive immune system of the ancient eukaryote, and have evolved into the cell-autonomous immune system that we know today. If this hypothesis is correct, future studies may find homologues of additional components of the human immune system functioning as phage resistance systems in bacteria.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1605-5>.

- Kranzusch, P. J. et al. Ancient origin of cGAS-STING reveals mechanism of universal 2',3' cGAMP signaling. *Mol. Cell* **59**, 891–903 (2015).
- Margolis, S. R., Wilson, S. C. & Vance, R. E. Evolutionary origins of cGAS-STING signaling. *Trends Immunol.* **38**, 733–743 (2017).
- Sun, L., Wu, J., Du, F., Chen, X. & Chen, Z. J. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science* **339**, 786–791 (2013).
- Ablasser, A. et al. cGAS produces a 2'-5'-linked cyclic dinucleotide second messenger that activates STING. *Nature* **498**, 380–384 (2013).
- Ishikawa, H., Ma, Z. & Barber, G. N. STING regulates intracellular DNA-mediated, type I interferon-dependent innate immunity. *Nature* **461**, 788–792 (2009).
- Davies, B. W., Bogard, R. W., Young, T. S. & Mekalanos, J. J. Coordinated regulation of accessory genetic elements produces cyclic di-nucleotides for *V. cholerae* virulence. *Cell* **149**, 358–370 (2012).
- Severin, G. B. et al. Direct activation of a phospholipase by cyclic GMP-AMP in *El Tor Vibrio cholerae*. *Proc. Natl Acad. Sci. USA* **115**, E6048–E6055 (2018).
- Makarova, K. S., Wolf, Y. I., Snir, S. & Koonin, E. V. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.* **193**, 6039–6056 (2011).
- Goldfarb, T. et al. BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* **34**, 169–183 (2015).
- Ofir, G. et al. DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.* **3**, 90–98 (2018).
- Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
- Iyer, L. M., Burroughs, A. M. & Aravind, L. The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like  $\beta$ -grasp domains. *Genome Biol.* **7**, R60 (2006).
- Kato, K., Ishii, R., Hirano, S., Ishitani, R. & Nureki, O. Structural basis for the catalytic mechanism of DncV, bacterial homolog of cyclic GMP-AMP synthase. *Structure* **23**, 843–850 (2015).
- Molineux, I. J. Host-parasite interactions: recent developments in the genetics of abortive phage infections. *New Biol.* **3**, 230–236 (1991).
- Walker, J. T. & Walker, D. H. Mutations in coliphage P1 affecting host cell lysis. *J. Virol.* **35**, 519–530 (1980).
- Whiteley, A. T. et al. Bacterial cGAS-like enzymes synthesize diverse nucleotide signals. *Nature* **567**, 194–199 (2019).
- Snyder, L. Phage-exclusion enzymes: a bonanza of biochemical and cell biology reagents? *Mol. Microbiol.* **15**, 415–420 (1995).
- Burroughs, A. M., Zhang, D., Schäffer, D. E., Iyer, L. M. & Aravind, L. Comparative genomic analyses reveal a vast, novel network of nucleotide-centric systems in biological conflicts, immunity and signaling. *Nucleic Acids Res.* **43**, 10633–10654 (2015).
- Joshua-Tor, L. & Hannon, G. J. Ancestral roles of small RNAs: an Ago-centric perspective. *Cold Spring Harb. Perspect. Biol.* **3**, a003772 (2011).
- Swarts, D. C. et al. DNA-guided DNA interference by a prokaryotic argonaute. *Nature* **507**, 258–261 (2014).
- Olovnikov, I., Chan, K., Sachidanandam, R., Newman, D. K. & Aravin, A. A. Bacterial argonaute samples the transcriptome to identify foreign DNA. *Mol. Cell* **51**, 594–605 (2013).
- Akira, S. & Takeda, K. Toll-like receptor signalling. *Nat. Rev. Immunol.* **4**, 499–511 (2004).
- Zhou, A. et al. Interferon action and apoptosis are defective in mice devoid of 2',5'-oligoadenylate-dependent RNase L. *EMBO J.* **16**, 6355–6363 (1997).
- Kazlauskienė, M., Kostiuik, G., Venclovas, Č., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* **357**, 605–609 (2017).
- Margulis, L. Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. *Proc. Natl Acad. Sci. USA* **93**, 1071–1076 (1996).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019



# Article

## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

### Genomic identification and analysis of DncV homologues

The protein sequence of DncV (NCBI accession NP\_229836) was searched against the protein sequences of all genes in 38,167 bacterial and archaeal genomes downloaded from the Integrated Microbial Genomes (IMG) database<sup>26</sup> in October 2017, using the 'search' option in the MMseqs2 package<sup>27</sup> (release 6-f5a1c) with default parameters. Hits with an *e* value less than  $1 \times 10^{-20}$  were taken as homologues. The fraction of homologues found in the vicinity of known defence systems was calculated as previously described<sup>11</sup>, using a positive set of defence gene families that was updated to include a recently discovered set of defence genes<sup>11</sup>.

### Genomic analysis of oligonucleotide cyclase genes

For the analysis in Fig. 3, the proteins from the database of 38,167 bacterial and archaeal genomes were first clustered using the 'cluster' option of MMseqs2<sup>27</sup> (release 2-1c7a89), with default parameters. Clusters were further aggregated into larger clusters using four additional cycles of clustering, in which—in each cycle—a representative sequence was taken from each cluster using the 'createsubdb' option of MMseqs2 and representative sequences were clustered using the 'cluster' option with the '-add-self-matches' parameter. For the first additional clustering cycle, the 'cluster' option was run with default parameters; for the additional cycles 2–4, clustering was run with sensitivity parameter '-s 7.5', and for the additional cycle 4, the '-cluster-mode 1' parameter was also added.

The sequences of each cluster were aligned using Clustal Omega<sup>28</sup>. Each multiple sequence alignment was scanned with HHpred<sup>29</sup> using 60% gap rule (-M 60) against the PDB\_mmcif<sup>30</sup> and pfam31<sup>31</sup> databases. Clusters with HHpred hits to one of the cGAS entries (Protein Data Bank (PDB) codes: 4LEV, 4MKP, 4O67, 5VDR, 5V8H, 4LEW, 5VDP, 4KM5, 4O68, 4O69, 5V8J, 5V8N, 5V8O, 5VDO, 5VDQ, 5VDS, 5VDT, 5VDU, 5VDV, 5VDW, 4XJ5, 4XJ1, 4XJ6, 4XJ3 and 4XJ4, and pfam PF03281) with >90% probability in the top 30 hits were taken for manual analysis. Clusters containing genes that were suspected to belong to toxin–antitoxin gene pairs were discarded. Overall, this procedure identified 30 clusters containing 6,232 predicted oligonucleotide cyclase genes from 5,150 genomes.

The genomic environments spanning 10 genes upstream and downstream of each of the 6,232 predicted oligonucleotide cyclase genes were searched to identify conserved gene cassettes and known defence genes around the oligonucleotide cyclase genes, as previously described<sup>11</sup>. Predicted systems were manually reviewed and unrelated genes (for example, mobilome genes and genes of other defence systems) were omitted.

To generate the phylogenetic tree in Fig. 3a, the 'clusthash' option of MMseqs2 (release 6-f5a1c) was first used to remove protein redundancies (using the '-min-seq-id 0.9' parameter). Sequences shorter than 200 amino acids were also removed. To further remove outlier sequences, an all-versus-all search was conducted using the 'search' option in MMseqs2, and proteins with less than 20 hits were manually examined. Overall, 17 outlier proteins were removed this way. The human cGAS protein (UniProt Q8N884) was added, as well as the human oligoadenylate synthase genes (UniProt P00973, P29728 and Q9Y6K5); these were used as an outgroup. Sequences were aligned using MAFFT<sup>28</sup> with gap open penalty of '-op 2'. The FastTree software<sup>32</sup> was used to generate a tree from the multiple sequence alignment using default parameters. The iTOL<sup>33</sup> software was used for tree visualization.

### Cloning of the CBASS into *E. coli* MG1655

The following full-length constructs containing the CBASS were designed: the four-gene CBASS from *V. cholerae* El Tor N16961 (GenBank accession NC\_002505) was identified as locus tags VCO178–VCO181,

and the operon was taken together with its upstream and downstream intergenic regions, spanning the nucleotide range 178,424–183,957 in GenBank accession NC\_002505. The four-gene system from *E. coli* TW11681 (GenBank accession AELD00000000) was identified as locus tags ESGDRAFT\_00026–ESGDRAFT\_00029 and the operon, together with its upstream and downstream intergenic regions, spanned the nucleotide range 21,738–27,072 in GenBank accession AELD00000000. These two constructs were commercially synthesized and cloned by Genscript directly into the plasmid pSG1-rfp<sup>10,11</sup> between the *Ascl* and *NotI* sites of the multiple cloning site.

For strains with gene deletions and point mutations, plasmids containing systems with these deletions or mutations were also commercially synthesized and cloned into the plasmid pSG1-rfp by Genscript, except for one construct in which the genes for both the E1–E2 domains and JAB domain were deleted (used for Extended Data Figs. 4, 5b). To build this construct, the pSG1-rfp backbone was amplified with primers P1-Fw + P1-Rv using KAPA HiFi HotStart ReadyMix (Kapa Biosystems KK2601) (Supplementary Table 2). Primers P2-Fw + P2-Rv were used to lift the gene pair *capV* and *dncV* with their native promoter from the plasmid that contained the full system derived from *E. coli* TW11681 (Supplementary Table 2). After gel purification of the PCR products with Zymoclean Gel DNA Recovery Kit (cat. no. D4001), the two fragments were assembled using NEBuilder HiFi DNA Assembly cloning kit (NEB E5520S) to produce the construct that lacks both gene C (encoding the protein with the E1 and E2 domains) and gene D (encoding the JAB-domain protein).

The plasmids were transformed into *E. coli* MG1655 cells by electroporation, and the resulting transformants of the wild-type and mutated systems were verified by whole genome sequencing as previously described<sup>10</sup> to verify system integrity and a lack of mutations. A negative control was constructed as a transformant containing an empty pSG1 plasmid.

### Cloning of the CBASS from *B. cereus* into *B. subtilis* BEST7003

The two-gene CBASS from *B. cereus* VD146 (GenBank accession KB976672) was identified as locus tags IK1\_05630–IK1\_05631, and the operon was taken together with its upstream and downstream intergenic regions, spanning the nucleotide range 60,974–63,493 in GenBank accession KB976672. The construct, as well as two additional constructs in which one of the two genes was deleted, were commercially synthesized and cloned by Genscript directly into the plasmid pSG1-rfp<sup>10,11</sup> between the *Ascl* and *NotI* sites of the multiple cloning site. The plasmid was transformed into *B. subtilis* and integrated into the *amyE* locus as previously described<sup>11</sup>. A negative control was constructed as a transformant containing an empty pSG1 plasmid integrated in the *amyE* locus. The resulting transformants of the wild-type and mutated systems were verified by whole genome sequencing as previously described<sup>10</sup>.

### Phage cultivation

*E. coli* phages (P1, T4, T5, T7 and  $\lambda$ -vir) were provided by U. Qimron. Phages SECphi17, SECphi18 and SECphi27 were isolated in our laboratory<sup>11</sup>. T2 and T6 were ordered from the Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ) (DSM 16352 and DSM 4622, respectively). The following *B. subtilis* phages were obtained from the *Bacillus* Genetic Stock Center (BGSC): SPO1 (BGSCID IP4),  $\phi$ 3T (BGSCID 1L1), SP $\beta$  (BGSCID 1L5), SPR (BGSCID 1L56),  $\phi$ 105 (BGSCID 1L11),  $\rho$ 14 (BGSCID 1L15), SPP1 (BGSCID IP7), SP82G (BGSCID IP5). Phage  $\phi$ 29 was obtained from the DSMZ (DSM 5546). Phages SBSphiJ and SBSphiC were isolated in our laboratory<sup>11</sup>. Phages were propagated on *E. coli* MG1655 or *B. subtilis* BEST7003 in liquid culture, and their titre was determined using the small drop plaque assay method, as previously described<sup>11</sup>.

### Plaque assays

Bacteria were mixed with MMB agar (LB + 0.1 mM MnCl<sub>2</sub> + 5 mM MgCl<sub>2</sub> + 0.5% agar), and tenfold serial dilutions of the phage lysate in MMB were dropped on top of them. After the drops were dry, plates were



incubated overnight at 37 °C for *E. coli* phage and at room temperature for *B. subtilis* phages. Plaques were counted to calculate the efficiency of plating in plaque-forming units per millilitre. For phages showing a fuzzy killing zone in which single plaques could not be counted, the lowest phage concentration in which a killing zone was observed was counted as ten plaques. Fold defence was calculated as the efficiency of plating on control bacteria divided by the efficiency of plating value obtained on bacteria containing the CBASS.

### Phage-infection dynamics in liquid medium

Overnight cultures were diluted 1:100 in MMB medium and incubated at 37 °C while shaking at 250 r.p.m. until early log phase ( $OD_{600} = 0.3$ ). One hundred and eighty microlitres of the diluted culture were transferred into wells in a 96-well plate containing 20 µl of phage lysate for a final MOI of 2, 0.2 or 0.02 as applicable. Infections were performed in triplicate and  $OD_{600}$  was followed using a TECAN Infinite 200 plate reader with measurement every 5 min.

### Transformation efficiency assay

To prepare electro-competent cells, *E. coli* MG1655 cells (with or without the CBASS) were diluted 1:100 in 100 ml LB medium supplemented with ampicillin, 100 µg/ml. At  $OD_{600} = 0.6$ , the cells were transferred to ice for 15 min and then centrifuged for 15 min at 4,000 r.p.m. The supernatant was discarded and the pellet was resuspended in ultrapure ice-cold water. This step was repeated twice and the supernatant was removed. The resultant pellet was resuspended in 10% glycerol, centrifuged for another 10 min and the supernatant was discarded. The final pellet was mixed in 500 µl of 10% glycerol and aliquots of 50 µl were flash-frozen in liquid nitrogen and transferred immediately to -80 °C.

One hundred nanograms of pRSFDuet-1 was added to 50 µl electro-competent cells and the mixture was transferred to a Bio-Rad Gene Pulser Curvette (0.2 cm, cat. no. 165-2086). The cells were electroporated with a Bio-Rad Micropulser using the 'Ec1' setting and then immediately transferred to 1 ml LB medium to recover at 37 °C for 1 h. After incubation cells were diluted 1:100, and 100 µl were plated on LB plates containing ampicillin (100 µg/ml) and kanamycin (50 µg/ml), and incubated at 37 °C overnight. Transformation efficiency was calculated by dividing the number of transformants that grew on LB plates containing ampicillin and kanamycin by the live count grown on LB ampicillin (100 µg/ml) only.

### Cloning, expression and purification of CapV from *E. coli* TW11681

The CapV protein from *E. coli* TW11681 was PCR-amplified (primers P3-Fw + P3-Rv, and P4-Fw + P4-Rv), fused with a C-terminal His tag (linker plus His tag sequence: SerGly<sub>4</sub>His<sub>6</sub>) and cloned into pET28b. The plasmid was transformed into BL21(DE3)pLys cells and grown at 37 °C, 250 r.p.m., until induction (1 mM IPTG) at an  $OD_{600}$  of 0.6. After induction, bacterial cell culture growth was continued for an additional 14 h at a temperature of 18 °C. Cells were then centrifuged for 10 min (3,900g, 4 °C) and pellets kept at -20 °C. Pellets were thawed on ice and resuspended with an ice-cold buffer containing 50 mM phosphate buffer (pH 7.4), 300 mM NaCl, 10% glycerol (v/v). The buffer was supplemented with 10 µg/ml lysozyme, and 1 µl benzonase-nuclease (SKU Merck). The suspension was then mixed with Lysing matrix B (MP) beads and cells were disrupted mechanically using a FastPrep-24 (MP) bead-beater device (2 cycles of 40 s, 6 m s<sup>-1</sup>). Cell lysate was centrifuged at 12,000g for 10 min at 4 °C and the supernatant was then mixed with Ni-NTA magnetic agarose beads (Qiagen) for 2 h (4 °C). CapV-6×His proteins bound to Ni-NTA beads were washed 3 times with a 50 mM phosphate buffer (pH 7.4), 300 mM NaCl, 10% glycerol (v/v), 20 mM imidazole and then eluted with the 50 mM phosphate buffer (pH 7.4), 300 mM NaCl, 10% glycerol (v/v), 250 mM imidazole. The eluent was loaded onto an Amicon Ultra-0.5 centrifugal filter unit 10-kDa filter (Merck) to exchange the elution buffer with the reaction buffer (50

mM phosphate buffer (pH 7.4), 300 mM NaCl, 10% glycerol (v/v)). Buffer exchange was done by centrifuging the centrifugal filter unit at 14,000g for 10 min at 4 °C 4 times with the reaction buffer. Eluted protein purification was assessed by running the sample on an SDS-PAGE gel and quantity was measured by using a Qubit Protein Assay Kit (Thermo Fisher Scientific).

### Fluorogenic biochemical assay for CapV activity

The esterase activity of the 6×His-tagged CapV was probed with the fluorogenic substrate resorufin butyrate. The 6×His-tagged CapV was diluted in 50 mM sodium phosphate (pH 7.4), 300 mM NaCl, 10% (v/v) glycerol to a final concentration of 1.77 µM. To measure the linear range of CapV activation, purified 6×His-tagged CapV was incubated for 5 min with increasing concentrations of 3'-cGAMP (Merck), ranging from 0.04 to 2.5 µM of 3'-cGAMP. In parallel, 6×His-tagged CapV was incubated for 5 min with cell lysates derived from *E. coli* cells infected with phage P1 or uninfected. Subsequently, the enzyme-cGAMP or enzyme-lysate solution was added to DMSO-solubilized resorufin butyrate (stock of 20 mM mixed with 50 mM sodium phosphate (pH 7.4), 300 mM NaCl, 10% (v/v) glycerol reaching a final concentration of 64 µM (final assay DMSO concentration was 0.32%)) to a final assay volume of 50 µl, and fluorescence was measured in a 96-well plate (Corning 96-well half area black non-treated plate with a flat bottom). Plates were read once every 30 s for 10 min at 37 °C using an Infinite-200 (Tecan) with excitation and emission wavelengths of 550 and 591 nm, respectively. The enzymatic reaction velocity was measured as previously described<sup>34</sup>. For this, a regression fit was calculated for the output values of each reaction over time, and the slope of this linear regression fit was used for determining the initial reaction velocity (FU s<sup>-1</sup>).

### Cell lysate preparation

*E. coli* MG1655 cells containing the *E. coli* TW11681-derived CBASS in which the *capV* gene was deleted were used for preparation of cell lysates. Cells containing an empty vector (pSG1) were used as control. Cells were grown in 100 ml MMB medium (flask size 250 ml) at 37 °C (250 r.p.m.) until reaching an  $OD_{600}$  of 0.3. Cells were then infected with 5 ml of phage P1 (titre of 10<sup>10</sup> infective particles per millilitre, estimated MOI of 2). After 30 or 40 min from the initial infection, samples were collected and centrifuged at 3,900g for 5 min at 4 °C. Following centrifugation, pellets were kept on ice until resuspended in 600 µl buffer containing 50 mM sodium phosphate (pH 7.4), 300 mM NaCl and 10% (v/v) glycerol. The resuspended pellet was supplemented with 1 µl hen-lysozyme (Merck) (final hen-lysozyme concentration of 16 µg/ml). The resuspended cells were then mixed with Lysing matrix B (MP) beads and cells were disrupted mechanically using a FastPrep-24 (MP) bead-beater device (2 cycles of 40 s, 6 m s<sup>-1</sup>, at 4 °C). Cell lysate was then centrifuged at 12,000g for 10 min at 4 °C and the supernatant was loaded onto a 3-kDa filter Amicon Ultra-0.5 centrifugal filter unit (Merck) and centrifuged at 14,000g for 30 min at 4 °C. The flow-through—containing substances smaller than 3 kDa—was used as the lysate sample for evaluating cGAMP production within the cells (Fig. 2).

### Quantification of 3'-cGAMP by high-performance liquid chromatography and mass spectrometry (HPLC-MS)

Cell lysates were prepared as described in 'Cell lysate preparation'. Lysates collected at 40 min after infection were analysed by MS-Omics using ultra-performance liquid chromatography (UPLC) (Vanquish, Thermo Fisher Scientific) coupled with a high-resolution quadrupole-orbitrap mass spectrometer (Q Exactive HF Hybrid Quadrupole-Orbitrap, Thermo Fisher Scientific). An electrospray ionization interface was used as an ionization source. Analysis was performed in positive ionization mode. A calibration series of 3'-cGAMP (SML1232, Sigma-Aldrich) ranging 0.001 to 50 µM was prepared and a linear regression from 0.001 to 5 µM was used for cGAMP quantification.

## FACS analysis of infected cells

Overnight cultures of *E. coli* MG1655 cells containing the *E. coli* TW11681-derived CBASS, and *E. coli* MG1655 cells containing an empty vector (pSG1), were diluted 1:100 in 0.5 ml MMB and grown at 37 °C and 500 rpm to an OD<sub>600</sub> of 0.3. Cells were then infected with phage P1 (titre of 10<sup>10</sup> infective particles per millilitre, estimated MOI of 2). At 40 min after infection, 40 µl of each culture were diluted into 2 ml of filtered PBS containing 1 µl of propidium iodide (Invitrogen LIVE/DEAD BacLight Bacterial Viability Kit (L7007)). The diluted bacteria were incubated in the dark in room temperature for five minutes and were then analysed by a BIORAD ZE5 Cell Analyzer. A 5-s agitation was performed and then 25,000 ungated events were recorded for both the CBASS-containing and CBASS-lacking cultures at 0.2 µl s<sup>-1</sup>. Forward scatter and propidium iodide fluorescence were measured using the height (*H*) parameter. FlowJo v.10 was used to analyse and visualize the data.

## Microscopy of infected cells

*E. coli* MG1655 cells that contain the *E. coli* TW11681-derived CBASS or the same CBASS with a point mutation of the phospholipase catalytic site (S60A) were grown in MMB medium at 37 °C. When growth reached an OD<sub>600</sub> of 0.3, bacteria were infected with P1 phage (MOI of about 2). Five hundred microlitres of the infected samples were centrifuged at 10,000g for 2 min at 25 °C and resuspended in 5 µl of 1× phosphate-buffered saline (PBS), supplemented with 1 µg/ml membrane stain FM4-64 (Thermo Fisher Scientific T-13320) and 2 µg/ml DNA stain 4,6-diamidino-2-phenylindole (DAPI) (Sigma-Aldrich D9542-5MG). Cells were visualized and photographed using an Axioplan2 microscope (ZEISS) equipped with ORCA Flash 4.0 camera (HAMAMATSU). System control and image processing were carried out using Zen software version 2.0 (Zeiss).

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

Data that support the findings of this study are available within the article and its Extended Data and Supplementary Tables. GenBank accessions, locus tags and nucleotide ranges of the CBASSs appear in the Methods. IMG gene and genome ID number, contig ID and system and effector

classification appear in Supplementary Table 1. Primer sequences for the CBASSs are available in Supplementary Table 2. Any other relevant data are available from the corresponding authors upon reasonable request.

26. Chen, I. A. et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).
27. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
28. Madeira, F. et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).
29. Zimmermann, L. et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
30. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003).
31. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
32. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
33. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
34. Lam, V. et al. Resorufin butyrate as a soluble and monomeric high-throughput substrate for a triglyceride lipase. *J. Biomol. Screen.* **17**, 245–251 (2012).
35. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols* **10**, 845–858 (2015).

**Acknowledgements** We thank A. Leavitt and S. Sharir for assistance in DNA extraction, library preparation and sequencing, A. Bernheim for assistance in data visualization, and members of the Sorek laboratory for fruitful discussions. This study was supported in part by the Israel Science Foundation (personal grant 1360/16), the European Research Council (grant ERC-CoG 681203), the Ernest and Bonnie Beutler Research Program of Excellence in Genomic Medicine, and the Knell Family Center for Microbiology. A.M. was supported by a fellowship from the Ariane de Rothschild Women Doctoral Program.

**Author contributions** D.C., S.M. and G.A. led the study and performed all experiments unless otherwise indicated. A.M. performed the computational analyses that appear in Figs. 1 and 3. Y.O.-S. performed the microscopy analysis that appears in Extended Data Fig. 7. G.S. assisted with the plaque assays that appear in Fig. 1 and Extended Data Figs. 1 and 3. A.K. and S.D. performed the computational analyses that led to Extended Data Fig. 8. R.S. supervised the study and wrote the paper together with the team.

**Competing interests** R.S. is a scientific cofounder and consultant of BiomX Ltd, Pantheon Ltd and Ecophage Ltd.

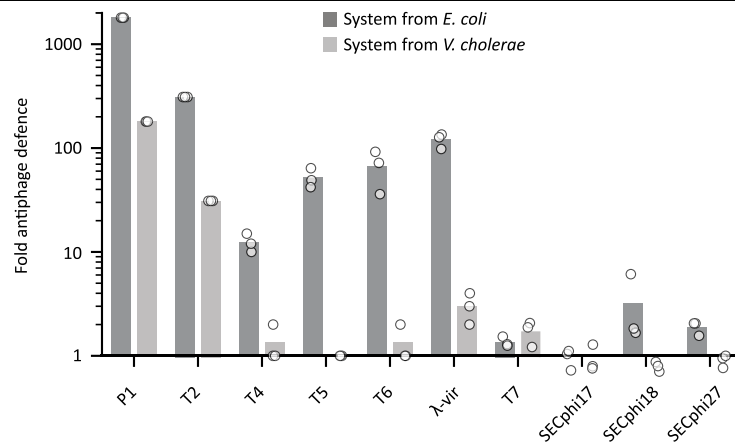
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1605-5>.

**Correspondence and requests for materials** should be addressed to G.A. or R.S.

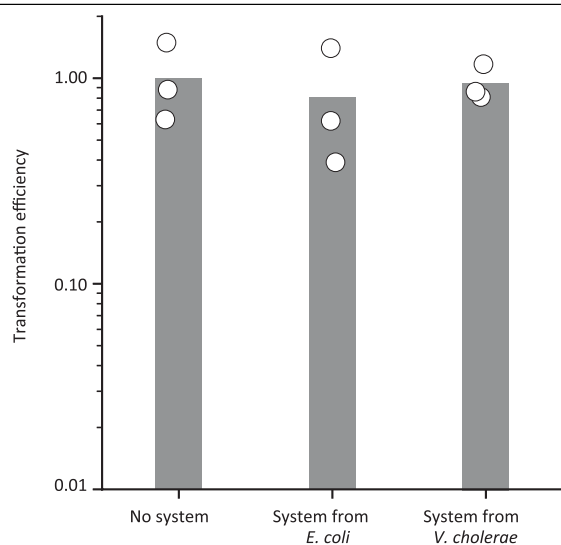
**Peer review information** Nature thanks Zhijian ‘James’ Chen, Karen Maxwell and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

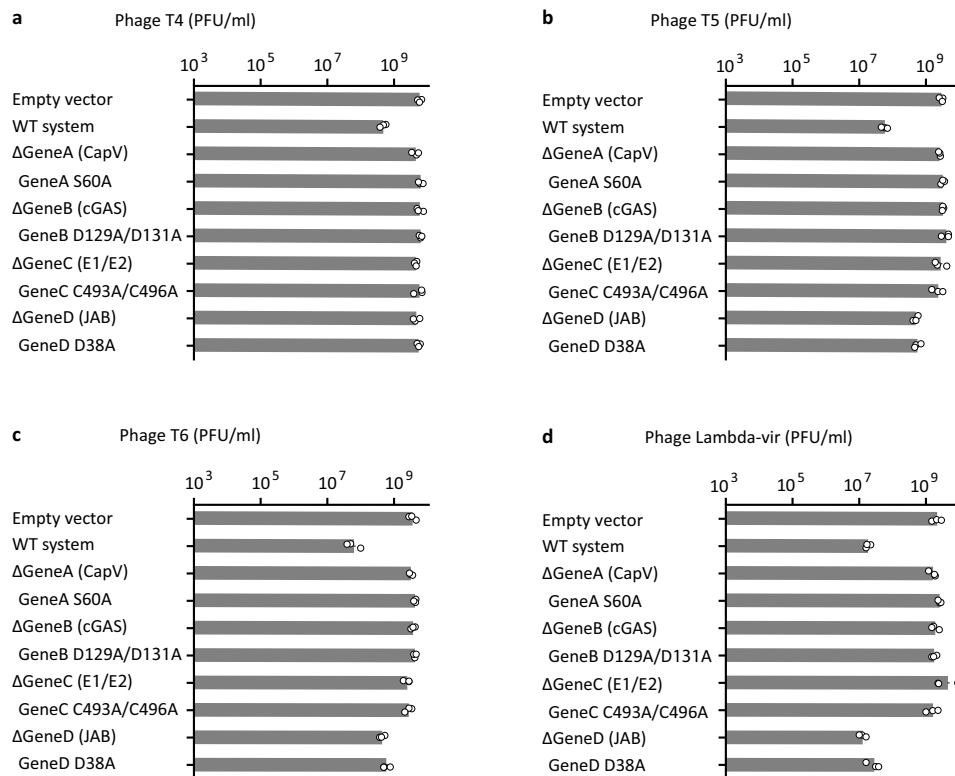


**Extended Data Fig. 1 | Fold antiphage defence conferred by four-gene defence systems against various phages.** The four-gene operon from either *V. cholerae* El Tor or *E. coli* TW11681 was cloned into *E. coli* MG1655 (Methods). Fold antiphage defence, as measured by plaque assays, is shown. The fold defence was calculated as the ratio between the efficiency of plating of the

phage on the operon-lacking control strain and the efficiency of plating on the operon-containing strain (Fig. 2b, Methods). Bar graph represents average of three independent replicates, with individual data points overlaid. Points that fall below the x axis (for SECphi17, SECphi18 and SECphi27) denote values lower than 1.



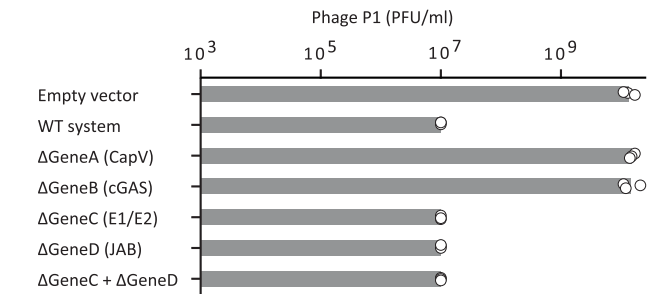
**Extended Data Fig. 2 | Transformation efficiency assays.** Transformation efficiency of plasmid pRSFDuet-1 into strains that contain the four-gene operon derived from *E. coli* TW11681 or from *V. cholerae* El Tor, presented as a fraction of the transformation efficiency to *E. coli* MG1655 carrying an empty vector instead of the four-gene operon. Bar graph represents average of three independent replicates, with individual data points overlaid.



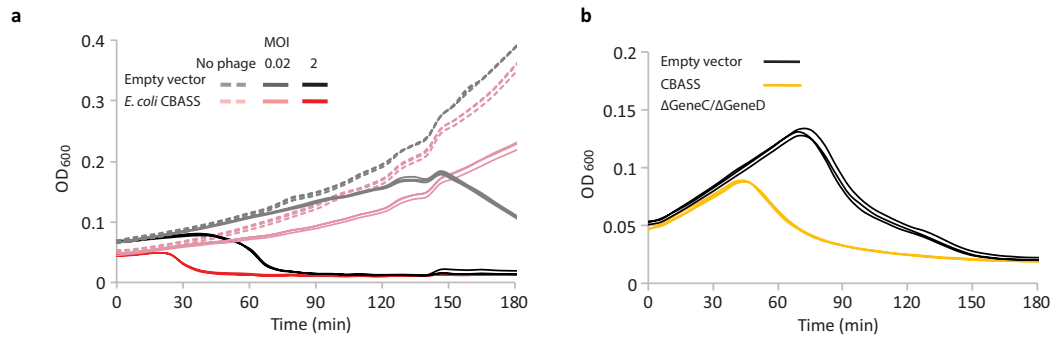
**Extended Data Fig. 3 | Efficiency of plating of coliphages on defence systems with whole-gene deletions or point mutations.** The efficiency of plating of phages infecting strains with the wild-type *E. coli*-derived four-gene, deletion strains and strains with point mutations. Data represent plaque-forming units per millilitre; bar graphs represent average of three independent replicates, with

individual data points overlaid. Empty vector represents a control *E. coli* MG1655 strain that lacks the system and has an empty vector instead. **a**, Infection with the phage T4. **b**, Infection with the phage T5. **c**, Infection with the phage T6. **d**, Infection with the phage λ-vir.



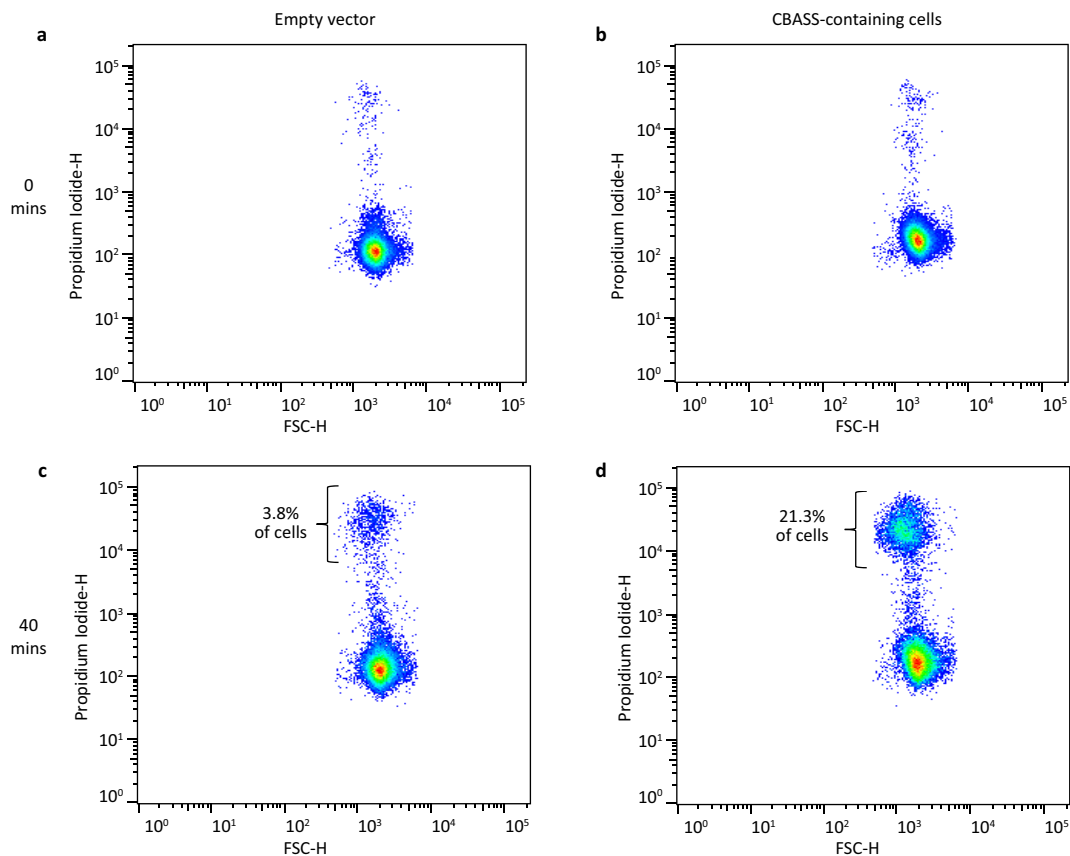


**Extended Data Fig. 4 | Efficiency of plating of phage P1 on a double-deletion strain.** The efficiency of plating is shown of phage P1 infecting strains with the wild-type *E. coli*-derived four-gene system, strains with individual genes deleted and a strain with two genes deleted. Data represent plaque-forming units per millilitre; bar graphs represent average of three independent replicates, with individual data points overlaid. Empty vector represents a control *E. coli* MG1655 strain that lacks the system and has an empty vector instead.



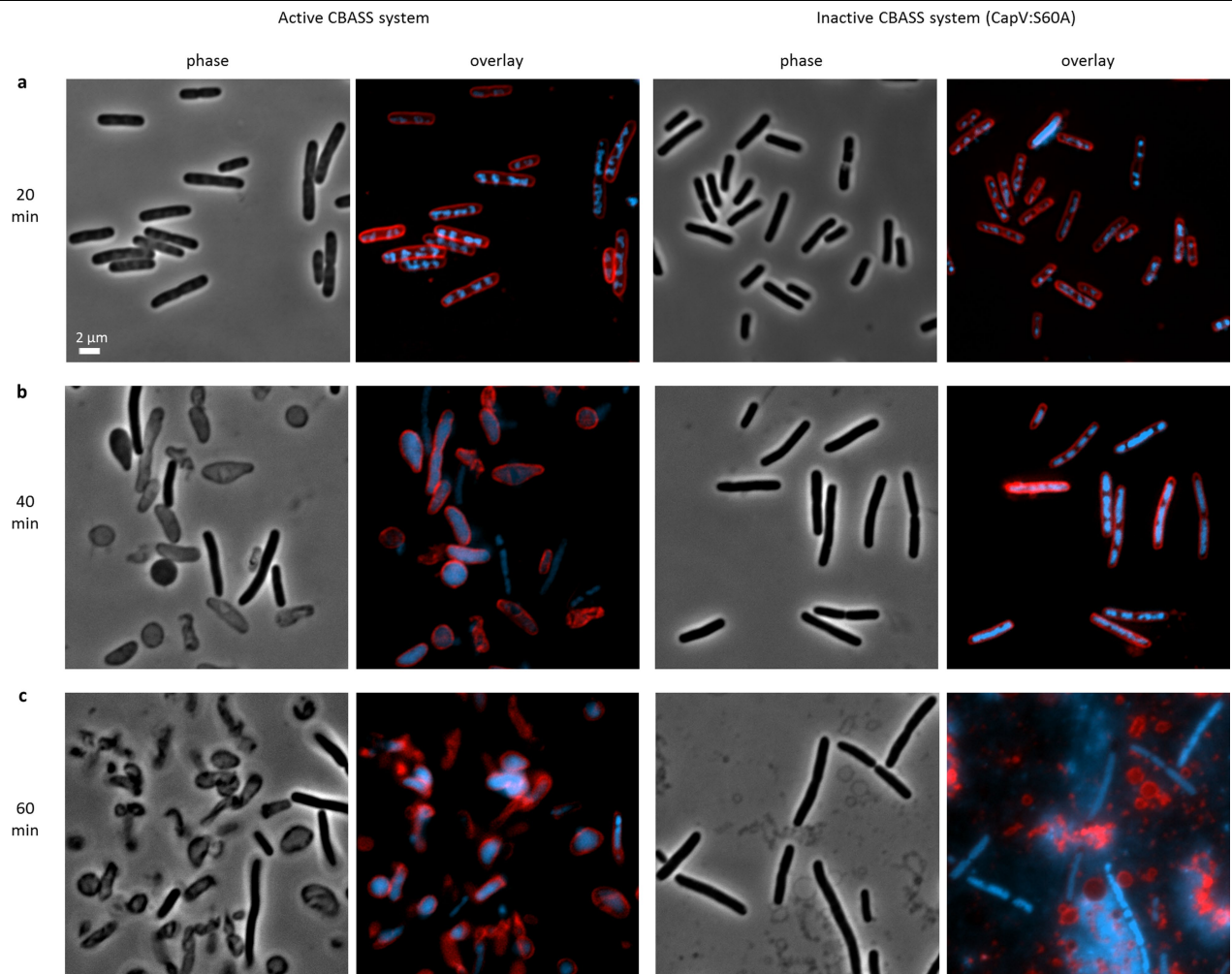
**Extended Data Fig. 5 | The bacterial CBASS functions through abortive infection. a,** Growth curves in liquid culture for CBASS-containing and CBASS-lacking (empty vector) bacteria infected by phage SECphi18 at 25 °C. Bacteria were infected at time = 0 at an MOI of 0.02 or 2. Three independent replicates for each MOI are shown, and each curve shows an individual replicate. **b,** Growth

curves in liquid culture for cells containing a minimal CBASS comprising phospholipase-cGAS (*capV-dncV*) only. Bacteria were infected at time = 0 at an MOI of 2 by phage P1. Three independent replicates for each MOI are shown, and each curve shows an individual replicate.



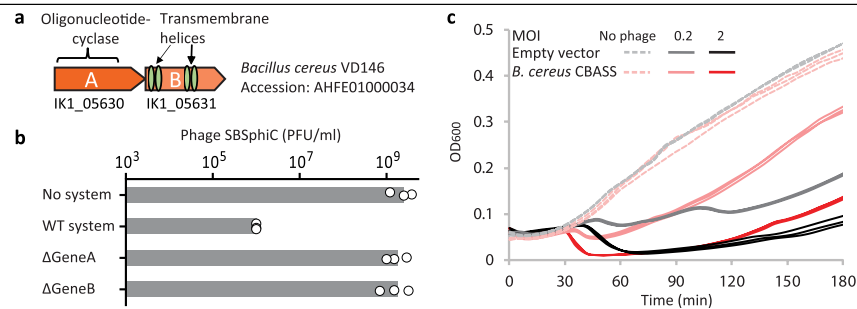
**Extended Data Fig. 6 | Cell sorting of infected cells stained with propidium iodide.** Cells containing the CBASS derived from *E. coli* TW11681, and control cells containing an empty vector, were stained with propidium iodide, a fluorescent DNA-binding agent that penetrates cells that have impaired membrane integrity. Cells were infected by phage P1 (MOI of 2) and sorted on the basis of propidium-iodide fluorescence intensity (y axis); the x axis represents

forward scatter. **a**, Uninfected cells that lack the CBASS. **b**, Uninfected cells that contain the CBASS. **c**, Cells that lack the CBASS, 40 min after infection. **d**, Cells that contain the CBASS, 40 min after infection. A large population of cells with high propidium-iodide fluorescence intensity is observed. Data from a representative replicate of two independent replicates are shown.



**Extended Data Fig. 7 | Microscopy of infected cells. a–c,** Phase contrast and overlay images are shown, of membrane stain (red) and DAPI (blue) images captured at 20 min (**a**), 40 min (**b**) and 60 min (**c**) after infection with phage P1 at an MOI of 2. The two columns on the left show *E. coli* MG1655 cells containing the CBASS derived from *E. coli* TW11681. The two columns on the right show *E. coli* MG1655 cells containing the CBASS derived from *E. coli* with a single point

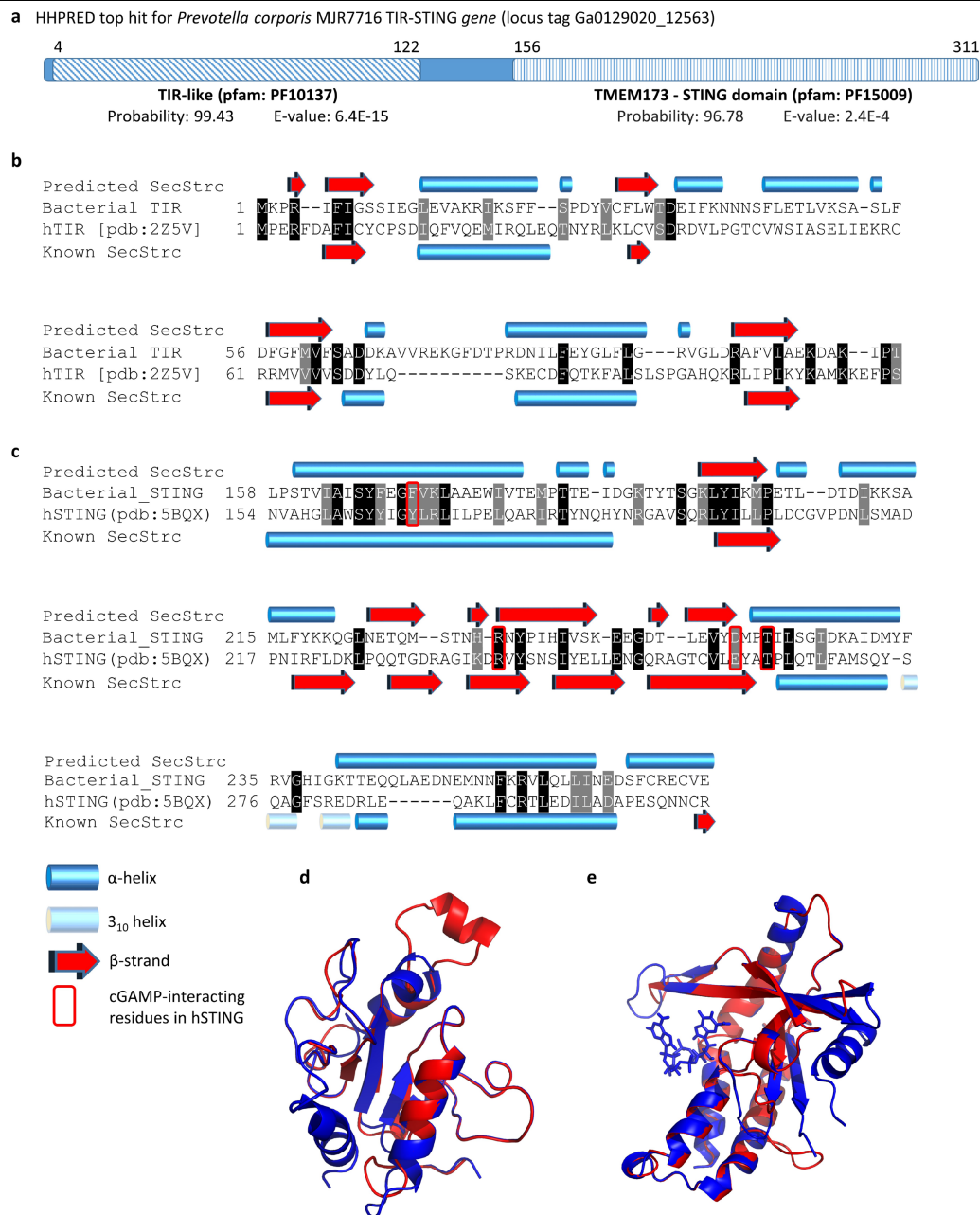
mutation that inactivates the CapV phospholipase (CapV(S60A)). Cell shape is deformed after 40 min in cells containing the CBASS, but not in cells in which the CBASS is mutated. After 60 min, phage-mediated cell lysis is observed in cells in which the CBASS is mutated. Representative images from a single replicate out of two independent replicates are shown.



**Extended Data Fig. 8 | A two-gene CBASS protects *Bacillus* against phage infection.** **a**, Domain organization of a two-gene operon found in the *B. cereus* VD146 genome. Locus tags of the depicted genes are indicated below each gene. **b**, The two-gene operon from *B. cereus* VD146 was cloned and genomically integrated into *B. subtilis* BEST7003, which naturally lacks this system. The efficiency of plating of phage SBSphiC infecting the CBASS-lacking and CBASS-containing strains, as well as strains in which one of the two genes was deleted, is

shown. Bar graph represents average of three independent replicates, with individual data points overlaid. **c**, Growth curves in liquid culture for *B. subtilis* containing the *B. cereus* two-gene CBASS, or CBASS-lacking *B. subtilis* that contains an empty vector instead, infected by phage SBSphiC. Bacteria were infected at time = 0 at an MOI of 0.2 or 2. Three independent replicates for each MOI are shown, and each curve shows an individual replicate.





**Extended Data Fig. 9 | Domain analysis and homology-based structure prediction of a bacterial TIR-STING protein.** **a**, Schematics of HHpred<sup>29</sup> homology-based search results of the *Prevotella corporis* TIR-STING protein (Supplementary Table 1). **b**, Phyre2<sup>35</sup> secondary structure prediction of the TIR domain in the *P. corporis* TIR-STING protein, compared to the solved crystal structure of the human TIR domain protein MyD88 (PDB accession 2Z5V\_A). **c**, Phyre2<sup>35</sup> secondary structure prediction of the STING domain in the *P. corporis* TIR-STING protein, compared to the solved crystal structure of the human

STING protein (PDB accession 5BQX\_A). Black, identical residues; grey, similar residues. Secondary structure prediction for the bacterial protein appears above the alignment; secondary structure of solved human domain appears below the alignment. **d**, Structural alignment of human TIR domain protein MYD88 and the modelled bacterial TIR domain. **e**, Structural alignment of human STING domain and the modelled bacterial STING domain. In **d**, **e**, blue and red represent the structure of the human protein and the model of the bacterial domain structure, respectively.

# MHC-II neoantigens shape tumour immunity and response to immunotherapy

<https://doi.org/10.1038/s41586-019-1671-8>

Received: 13 February 2019

Accepted: 12 September 2019

Published online: 23 October 2019

Elise Alspach<sup>1,2</sup>, Danielle M. Lussier<sup>1,2</sup>, Alexander P. Miceli<sup>1,2</sup>, Ilya Kizhvatov<sup>1</sup>, Michel DuPage<sup>3,8</sup>, Adrienne M. Luoma<sup>4</sup>, Wei Meng<sup>1,2</sup>, Cheryl F. Lichti<sup>1,2</sup>, Ekaterina Esaulova<sup>1</sup>, Anthony N. Vomund<sup>1</sup>, Daniele Runci<sup>1,2</sup>, Jeffrey P. Ward<sup>1,2,5</sup>, Matthew M. Gubin<sup>1,2</sup>, Ruan F. V. Medrano<sup>1,2</sup>, Cora D. Arthur<sup>1,2</sup>, J. Michael White<sup>1</sup>, Kathleen C. F. Sheehan<sup>1,2</sup>, Alex Chen<sup>1</sup>, Kai W. Wucherpfennig<sup>4</sup>, Tyler Jacks<sup>3,6</sup>, Emil R. Unanue<sup>1</sup>, Maxim N. Artyomov<sup>1</sup> & Robert D. Schreiber<sup>1,2,7\*</sup>

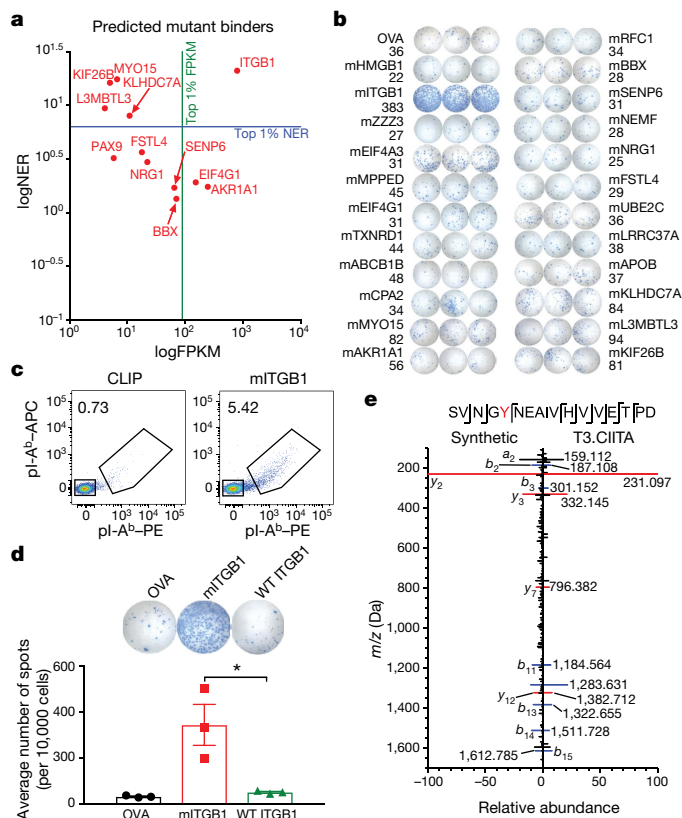
The ability of the immune system to eliminate and shape the immunogenicity of tumours defines the process of cancer immunoediting<sup>1</sup>. Immunotherapies such as those that target immune checkpoint molecules can be used to augment immune-mediated elimination of tumours and have resulted in durable responses in patients with cancer that did not respond to previous treatments. However, only a subset of patients benefit from immunotherapy and more knowledge about what is required for successful treatment is needed<sup>2–4</sup>. Although the role of tumour neoantigen-specific CD8<sup>+</sup> T cells in tumour rejection is well established<sup>5–9</sup>, the roles of other subsets of T cells have received less attention. Here we show that spontaneous and immunotherapy-induced anti-tumour responses require the activity of both tumour-antigen-specific CD8<sup>+</sup> and CD4<sup>+</sup> T cells, even in tumours that do not express major histocompatibility complex (MHC) class II molecules. In addition, the expression of MHC class II-restricted antigens by tumour cells is required at the site of successful rejection, indicating that activation of CD4<sup>+</sup> T cells must also occur in the tumour microenvironment. These findings suggest that MHC class II-restricted neoantigens have a key function in the anti-tumour response that is nonoverlapping with that of MHC class I-restricted neoantigens and therefore needs to be considered when identifying patients who will most benefit from immunotherapy.

Immune checkpoint therapy (ICT) demonstrates remarkable clinical efficacy in subsets of patients with cancer, but many patients do not show durable responses<sup>2–4</sup>. Although MHC class I (MHC-I)-restricted neoantigens are important targets of tumour-specific CD8<sup>+</sup> cytotoxic T lymphocytes (CTLs) during successful ICT in both mice and humans<sup>5–12</sup>, current methods to predict patient response to ICT are imprecise and additional or better prognostic indicators are needed<sup>13–17</sup>. The influence of MHC class II (MHC-II)-restricted CD4<sup>+</sup> T cell responses to tumour neoantigens during immunotherapy has only recently been addressed<sup>18,19</sup>. While some reports show that effective tumour immunity can occur in the absence of help from CD4<sup>+</sup> T cells, most indicate that CD4<sup>+</sup> T cells are important for generating tumour-specific CD8<sup>+</sup> T cells<sup>20–25</sup>. However, as it has proven difficult to identify tumour-specific mutations that function as neoantigens for CD4<sup>+</sup> T cells using existing MHC-II antigen prediction algorithms, considerable uncertainty remains as to whether strict tumour specificity in the CD4<sup>+</sup> T cell compartment is required during spontaneous or ICT-induced anti-tumour responses<sup>24,26,27</sup>, especially for tumours that do not express MHC-II.

In this study, we used the well-characterized, MHC-II-negative T3 methylcholanthrene (MCA)-induced sarcoma line, which grows progressively in wild-type mice but is rejected following ICT in a CD4<sup>+</sup> and CD8<sup>+</sup> T cell-dependent manner<sup>9</sup>. Although we have identified point mutations in laminin- $\alpha$  subunit 4 (LAMA(G1254V); mLAMA4) and asparagine-linked glycosylation 8 glucosyltransferase (ALG8(A506T); mALG8) as major MHC-I neoantigens in T3 cells, the identities of T3-specific MHC-II antigens remain unknown<sup>9</sup>. Here we use new predictive algorithms to identify an N710Y somatic point mutation in integrin- $\beta$ 1 (mITGB1) as a major MHC-II neoantigen of T3 sarcoma cells. In nonimmunogenic oncogene-driven KP9025 sarcoma cells (KP), which lack mutational neoantigens, co-expression of single MHC-I and MHC-II T3 neoantigens rendered KP9025 cells susceptible to ICT. We find similar requirements for vaccines that drive rejection of T3 tumours. In mice bearing contralateral KP.mLAMA4.mITGB1 and KP.mLAMA4 tumours, ICT induced the rejection of tumours expressing both neoantigens but not tumours expressing mLAMA4 only, indicating that co-expression of both MHC-I and MHC-II neoantigens at the tumour site is necessary for successful

<sup>1</sup>Department of Pathology and Immunology, Washington University School of Medicine, St Louis, MO, USA. <sup>2</sup>The Andrew M. and Jane M. Bursky Center for Human Immunology and Immunotherapy Programs, Washington University School of Medicine, St Louis, MO, USA. <sup>3</sup>David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>4</sup>Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, MA, USA. <sup>5</sup>Division of Oncology, Department of Medicine, Washington University School of Medicine, St Louis, MO, USA. <sup>6</sup>Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>7</sup>The Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA. <sup>8</sup>Present address: Division of Immunology and Pathogenesis, Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, CA, USA.

\*e-mail: rdschreiber@wustl.edu



**Fig. 1 | mITGB1 is a major MHC-II-restricted neoantigen of T3 sarcoma cells.** **a**, hmMHC predictions of MHC-II neoantigens expressed in T3 sarcoma cells. Potential neoantigens were filtered as shown in Extended Data Fig. 3a and those that met the strong binder threshold are shown as expression level (fragments per kilobase of transcript per million mapped reads; FPKM) and neoepitope ratio (NER). Strong binders are those with  $-\log_{10}\text{Odds} \leq 26.21$ . Green line, high-expression cut-off (FPKM = 89.1); blue line, high NER cut-off (NER = 6.55). **b**, CD4<sup>+</sup> T cells isolated from T3 TILs 12 days after transplantation were stimulated in IFN $\gamma$  ELISPOT analysis with naive splenocytes pulsed with 2  $\mu\text{g ml}^{-1}$  of the indicated individual peptide. Numbers beneath peptides represent average number of spots from three independent experiments. **c**, I-A<sup>b</sup> tetramer staining of CD4<sup>+</sup> T cells from whole T3 TILs 12 days after transplantation. Cells were gated on viable CD11b<sup>+</sup>CD4<sup>+</sup> cells. Representative data from one of three independent experiments are shown. WT, wild-type. **d**, Freshly isolated CD4<sup>+</sup> T cells from day 12 TILs were stimulated with 2  $\mu\text{g ml}^{-1}$  mITGB1(710Y) or wild-type ITGB1(710N) peptide-pulsed splenocytes and analysed by IFN $\gamma$  ELISPOT. Data are mean  $\pm$  s.e.m. ( $n = 3$  independent experiments). \* $P = 0.03$  (unpaired, two-tailed  $t$ -test). **e**, Mirror plot showing match between MS/MS spectra of the 17-mer peptide encompassing mITGB1(N710Y) eluted from T3.CIITA cells (right) and a corresponding synthetic peptide (left). Labelled  $m/z$  values reflect those experimentally observed for the endogenous peptide, with peaks representing  $b$  ions in blue and  $y$  ions in red.

ICT. These results show that the expression of MHC-II neoantigens in tumours is a critical determinant of responsiveness to ICT, personalized cancer vaccines and potentially other immunotherapies.

## Predicting MHC-II neoantigens with hmMHC

The best currently available methods for predicting MHC-II-restricted neoantigens rely on tools (netMHCII-2.3 and netMHCIIpan-3.2) that are inaccurate, partially because the open structure of the MHC-II binding groove leads to substantial variation in epitope length<sup>18,26</sup>. Moreover, the existing tools cannot be re-trained on new data. We therefore developed a hidden Markov model (HMM)-based MHC binding predictor (hmMHC,

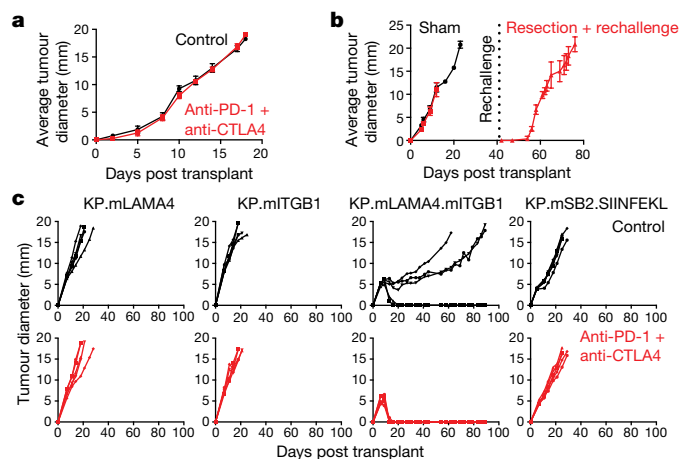
Extended Data Fig. 1a) that inherently accommodates peptide sequences of variable length and is trained on recent Immune Epitope Database (IEDB) content (Extended Data Fig. 1b–d). Validation analyses showed that hmMHC displays substantially higher sensitivity for high-specificity values than other predictors (Extended Data Fig. 2a, b). Using hmMHC, we calculated the likelihood of each of the 700 missense mutations that are expressed in T3 (Supplementary Data 1) being presented by the MHC-II I-A<sup>b</sup> molecule and refined our results by prioritizing candidates based on I-A<sup>b</sup> binding affinity, mutant:wild-type I-A<sup>b</sup> binding ratios, and transcript abundance<sup>18</sup> (Fig. 1a, Extended Data Fig. 3a).

One candidate, mITGB1, met all our criteria (Fig. 1a, Extended Data Fig. 3a). Notably, mITGB1 was not selected using netMHCII-2.3 or netMHCIIpan-3.2 (Extended Data Fig. 3b, data not shown). Enzyme-linked immune absorbent spot (ELISPOT) analysis showed that the mITGB1 peptide induced high IFN $\gamma$  production from CD4<sup>+</sup> T3 tumour-infiltrating lymphocytes (TILs). Other mutant peptides that fulfilled some but not all of our criteria induced only weak or absent responses, thereby validating our hmMHC prediction method (Fig. 1b, Extended Data Fig. 3c, Supplementary Table 1). To confirm this result, we stained T3-derived CD4<sup>+</sup> TILs with MHC-II tetramers carrying either the 707–721 mITGB1 peptide or an irrelevant peptide (CLIP). Whereas 5.9% of T3-infiltrating CD4<sup>+</sup> T cells stained positively with the mITGB1–I-A<sup>b</sup> tetramer, the CLIP–I-A<sup>b</sup> tetramer stained only 0.7% of the cells (Fig. 1c, Extended Data Fig. 3d, e). Cytokine profiling of mITGB1-specific CD4<sup>+</sup> TILs from T3 tumours revealed that they produced IFN $\gamma$ , TNF, and IL-2 but not IL-4, IL-10, IL-17 or IL-22, indicating a phenotype resembling that of T helper type 1 (T<sub>H</sub>1) cells (Extended Data Fig. 3f). T3 tumour-bearing mice treated with ICT did not develop additional MHC-II neoantigen specificities (data not shown). To assess whether T3-specific CD4<sup>+</sup> T cells selectively recognized the mutant, we compared mutant to wild-type ITGB1 peptides in ELISPOT analyses using freshly isolated T3 CD4<sup>+</sup> TILs. Only the mITGB1 peptide induced positive responses (Fig. 1d). Similar data were obtained using CD4<sup>+</sup> T cell hybridomas generated from T3 TILs (Extended Data Figs. 4, 5a).

Mapping experiments revealed that the MHC-II binding core of mITGB1 consists of nine amino acids (<sup>710</sup>YNEAIVHV<sup>718</sup>), in which the mutant Y710 residue functions as an I-A<sup>b</sup> anchor (Extended Data Fig. 5b). To verify that the mITGB1 epitope is physiologically presented by MHC-II, we transduced T3 cells with a vector encoding the mouse MHC-II transactivator CIITA (T3.CIITA cells), which induced high levels of I-A<sup>b</sup> expression<sup>28</sup> (Extended Data Fig. 5c). Elution of peptides bound to I-A<sup>b</sup> on T3.CIITA cells and analysis by mass spectrometry identified two mITGB1 peptides encompassing the Y710 mutation (a 17-mer and a 14-mer; Fig. 1e, Extended Data Fig. 5d). Peptides with the corresponding wild-type sequence were not found. The mITGB1 epitope was also not detected in MHC-I eluates from IFN $\gamma$ -stimulated T3 cells, and mITGB1-specific CD8<sup>+</sup> T cells were not observed by cytokine production (data not shown). Together, these data demonstrate that mITGB1 is a major MHC-II-restricted neoantigen of T3 sarcoma cells.

## ICT response requires CD4<sup>+</sup> T cell help

Recent publications have highlighted the ability of CD4<sup>+</sup> T cells to recognize tumour-specific antigens and promote tumour rejection in the absence of ICT<sup>18,29,30</sup>. To assess whether CD4<sup>+</sup> T cells are required during ICT-induced rejection, we expressed MHC-I and/or MHC-II neoantigens from T3 sarcoma cells in an oncogene-driven sarcoma cell line generated from a *Kras*<sup>LSL-G12D/+</sup> × *Tp53*<sup>R130C</sup> mouse injected intramuscularly with lentiviral Cre-recombinase (KP9025 cells)<sup>7</sup>. The unmodified KP9025 sarcoma line formed progressively growing tumours in either syngeneic wild-type mice treated with or without dual anti-PD-1 and anti-CTLA4 ICT or mice rechallenged with unmodified KP9025 after previously being cured of their KP9025 tumours via surgical resection (Fig. 2a, b). As this challenge–resection–rechallenge approach promotes immune control or rejection of even poorly immunogenic tumour cells used in



**Fig. 2 | ICT-mediated rejection of a nonimmunogenic sarcoma requires CD4<sup>+</sup> and CD8<sup>+</sup> T cells.** **a**, One million KP9025 sarcoma cells were injected subcutaneously into syngeneic 129S4 mice and animals were treated with either a control monoclonal antibody or the anti-PD-1 + anti-CTLA4 combination on days 3, 6, and 9 after transplantation. Representative data from two independent experiments are shown as mean tumour diameter  $\pm$  s.e.m. ( $n=5$  in all groups per experiment). **b**, KP9025 sarcoma cells were injected as above and tumours were surgically resected followed by rechallenge with the same line. Representative data from one of two independent experiments are shown as mean tumour diameter  $\pm$  s.e.m. ( $n=3$  in all groups per experiment). **c**, Cohorts of five mice were injected with  $1 \times 10^6$  KP.mLAMA4, KP.mITGB1, KP.mLAMA4.mITGB1, or KP.mSB2.SIINFELK cells and treated with either control monoclonal antibody (top) or the anti-PD-1 + anti-CTLA4 combination (bottom) on days 3, 6, and 9 after transplantation. Representative data from one of three independent experiments are shown as individual tumour diameters.

the initial priming step<sup>31</sup>, these results supported the conclusion that KP9025 sarcoma cells were not immunogenic. Whole-exome sequencing revealed that KP9025 cells expressed only four nonsynonymous mutations (Supplementary Data 2) and none were predicted to be immunogenic (Extended Data Fig. 6a, b, Supplementary Table 2). Enforced expression of either mLAMA4 or mITGB1 alone did not render KP9025 cells immunogenic in wild-type mice in the presence or absence of ICT (Fig. 2c, Extended Data Fig. 6d, e). Progressively growing KP.mLAMA4 tumours maintained expression of their MHC-I tumour neoantigen, thereby ruling out antigen loss via immunoediting (Extended Data Fig. 7a). KP9025 cells expressing both mLAMA4 and mITGB1 formed tumours in immunodeficient *Rag2*<sup>-/-</sup> mice that grew with kinetics similar to those of KP.mLAMA4 or KP.mITGB1 cells (Extended Data Fig. 6c). However, growth of KP.mLAMA4.mITGB1 cells in wild-type mice treated with a control monoclonal antibody was noticeably slower than that of either single-antigen-expressing cell line, and KP.mLAMA4.mITGB1 tumours were rejected in wild-type mice following either dual or single agent ICT despite the absence of tumour cell MHC-II expression (Fig. 2c, Extended Data Fig. 6d, e, data not shown).

We considered the possibility that the enhanced immunogenicity of KP.mLAMA4.mITGB1 tumours was merely a function of antigen quantity. Therefore, we generated KP9025 cells that lacked MHC-II neoantigens but co-expressed two strong MHC-I neoantigens: the MHC-I epitope of ovalbumin (SIINFELK) and the R913L mutant of spectrin- $\beta$ 2 (mSB2), which contributes to the spontaneous rejection of the MCA-induced d42m1 sarcoma line in wild-type mice<sup>6</sup>. KP.mSB2.SIINFELK tumours grew progressively in mice treated either with a control monoclonal antibody or dual ICT, and the expression of both MHC-I antigens was maintained in growing tumours from ICT-treated animals (Fig. 2c, Extended Data Fig. 7b–d). Enforced expression of mITGB1 in KP.mSB2.SIINFELK cells led to significantly ( $P=1.5 \times 10^{-5}$ ) increased survival of ICT-treated mice injected with the uncloned tumour line (Extended Data Fig. 7e). Thus,

tumour rejection and ICT sensitivity are dependent on combinatorial effects of CD4<sup>+</sup> and CD8<sup>+</sup> T cells.

### mITGB1 CD4<sup>+</sup> T cells are T<sub>H</sub>1 polarized

We next investigated whether mITGB1-specific CD4<sup>+</sup> TILs displayed a T<sub>H</sub>1 phenotype similar to that seen with T3 tumours. Seventy-four per cent of mITGB1 tetramer-positive CD4<sup>+</sup> T cells in KP.mLAMA4.mITGB1 tumours from control-treated mice expressed the T<sub>H</sub>1-associated transcription factor T-BET, but not the regulatory T cell (T<sub>reg</sub>)-associated transcription factor FOXP3. An additional 17% expressed both T-BET and FOXP3. Conversely, tetramer-negative CD4<sup>+</sup> T cells showed substantially diminished expression of T-BET (24%) and much higher expression of FOXP3 expression (61%). mITGB1-tetramer<sup>+</sup> CD4<sup>+</sup> T cells displayed a higher T-BET<sup>+</sup>:FOXP3<sup>+</sup> ratio than tetramer-negative cells (4 versus 0.4, respectively) and this ratio was further increased in response to anti-CTLA4 treatment (33 versus 3.7, respectively; Extended Data Fig. 8a–c). On average, 83% of mITGB1-specific CD4<sup>+</sup> T cells expressed high levels of PD-1 compared to only 19% of mITGB1-tetramer-negative cells (Extended Data Fig. 8d, e). CD4<sup>+</sup> T cells specific for mITGB1 also expressed high levels of CD44, ICOS and CD150 (also known as SLAMF6), and low levels of KLRG1 (Extended Data Fig. 8f). The presence of an expanded population of T<sub>H</sub>1-like ICOS<sup>+</sup> CD4<sup>+</sup> T cells was recently reported in mice bearing B16 or MC38 tumours that were treated with anti-CTLA4, although the tumour antigen specificity of this population was not identified<sup>32</sup>. These data, together with the cytokine profiles described above, indicate that mITGB1-specific CD4<sup>+</sup> T cells display an activated T<sub>H</sub>1 phenotype.

### CTL generation requires CD4<sup>+</sup> T cell help

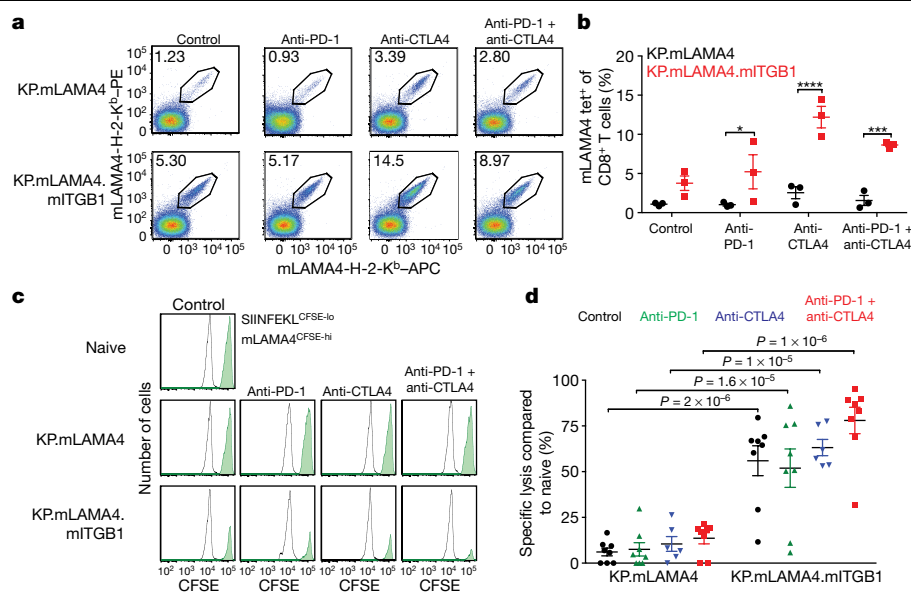
To identify the mechanism by which tumour neoantigen-specific CD4<sup>+</sup> T cells influence ICT-mediated anti-tumour responses, we assessed their effects on CD8<sup>+</sup> T cell priming by comparing MHC-I tetramer staining of splenic mLAMA4-specific CD8<sup>+</sup> T cells from mice bearing KP.mLAMA4 or KP.mLAMA4.mITGB1 tumours, treated with a control monoclonal antibody or ICT. In the absence of ICT, mLAMA4-H-2K<sup>b</sup> tetramers stained only 1.2% of CD8<sup>+</sup> T cells from mice bearing KP.mLAMA4 tumours, but 5.3% of CD8<sup>+</sup> T cells in mice bearing KP.mLAMA4.mITGB1 tumours (Fig. 3a, b). This staining percentage was unchanged in the presence of PD-1 blockade, but was increased by anti-CTLA4 treatment, either as monotherapy or in combination with anti-PD-1. This result is consistent with the observation that anti-CTLA4 treatment functions largely to enhance CD4<sup>+</sup> T cell responses<sup>32,33</sup>.

To assess whether MHC-II neoantigens also enhanced CTL formation, we used an in vivo T cell cytotoxicity assay that monitored the capacity of naturally arising CTLs to kill peptide-pulsed splenocytes labelled with carboxyfluorescein succinimidyl ester (CFSE)<sup>34</sup>. Non-tumour-bearing control mice and mice bearing KP.mLAMA4 tumours were largely incapable of eliminating mLAMA4 peptide-pulsed splenocytes in either the presence or absence of ICT (Fig. 3c). By contrast, mice bearing KP.mLAMA4.mITGB1 tumours efficiently eliminated CFSE<sup>hi</sup>-labelled, mLAMA4 peptide-pulsed splenocytes but not CFSE<sup>lo</sup>-labelled SIINFELK-pulsed splenocytes, and the degree of elimination of the former was enhanced by ICT (Fig. 3c, d). The cytotoxic activity of control-treated mLAMA4-specific CD8<sup>+</sup> T cells observed in the splenocyte killing assay was higher than would be expected from our in vivo tumour rejection experiments (Fig. 2e). This difference is likely to reflect differences in the susceptibility of splenocytes and tumour cells to T cell-mediated killing. Thus, CD4<sup>+</sup> T cell help enhances both CD8<sup>+</sup> T cell priming and maturation of CD8<sup>+</sup> T cells into CTLs.

### Vaccines require MHC-I and MHC-II antigens

As CD4<sup>+</sup> T cell help was crucial for generating mLAMA4-specific CTLs during ICT, we tested whether mITGB1-specific CD4<sup>+</sup> T cells were





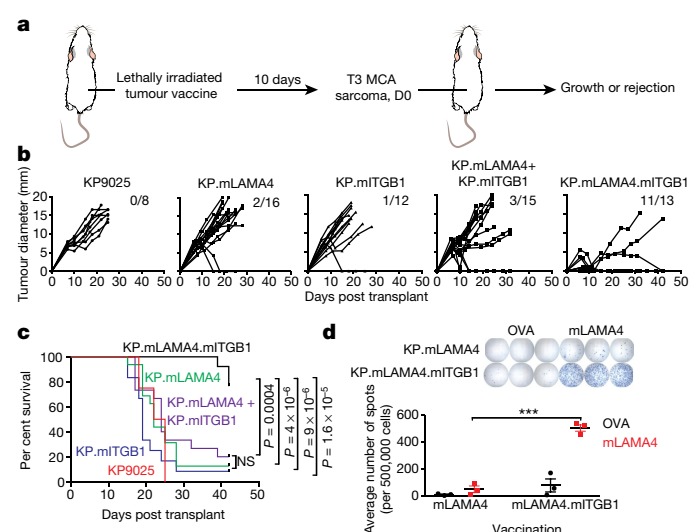
**Fig. 3 | CD4<sup>+</sup> T cell help is required for the generation of functional CD8<sup>+</sup> CTLs during ICT.** **a**, Representative tetramer staining of mLAMA4-specific CD8<sup>+</sup> T cells from the spleens of mice bearing KP.mLAMA4 (left) or KP.mLAMA4.mITGB1 (right) tumours 12 days after transplantation. Mice received the indicated ICT treatment on days 3, 6, and 9. Cells were gated from viable CD45<sup>+</sup>CD11b<sup>+</sup>Thy1.2<sup>+</sup> cells. **b**, Quantification of three independent experiments described in **a**, shown as mean  $\pm$  s.e.m. per cent mLAMA4 tetramer-positive CD8<sup>+</sup> T cells. \* $P=0.04$ , \*\*\* $P=0.0007$ , \*\*\*\* $P=0.00003$  (two-way ANOVA with multiple comparisons corrected with the Bonferroni method). **c**, In vivo cytotoxic function of mLAMA4-specific CD8<sup>+</sup> T cells. Naive splenocytes were labelled with

0.5  $\mu$ M CFSE and pulsed with 1  $\mu$ M SIINFEKL peptide (white histograms) or labelled with 5  $\mu$ M CFSE and pulsed with 1  $\mu$ M mLAMA4 peptide (green histograms) and transferred into control naive or tumour-bearing mice 11 days after tumour transplantation. Tumour-bearing mice received the indicated ICT treatment on days 3, 6, and 9 after transplantation. Representative data from six (anti-CTLA4) and eight (all other groups) independent experiments are shown. **d**, Quantification of per cent mLAMA4-specific lysis from independent in vivo cytotoxicity assays in **c** shown as mean  $\pm$  s.e.m. ( $n=6$  in anti-CTLA4,  $n=8$  in all other groups).  $P$  values calculated using two-way ANOVA with multiple comparisons and Bonferroni correction.

also important for vaccine-elicited anti-tumour responses (Fig. 4a). Vaccination of naive recipient mice with irradiated parental KP9025, KP.mLAMA4, or KP.mITGB1 cells was not sufficient to protect most mice from a subsequent challenge with T3 sarcoma cells. Vaccination with a mixture of irradiated KP.mLAMA4 and KP.mITGB1 cells provided protection against T3 challenge in 30% of mice. By contrast, vaccination with irradiated KP.mLAMA4.mITGB1 cells prevented T3 tumour outgrowth in 11 of 13 recipients (Fig. 4b, c). Furthermore, spleens from mice vaccinated with irradiated KP.mLAMA4.mITGB1 cells contained significantly ( $P=0.0002$ ) more mLAMA4-specific, IFN $\gamma$ -producing CD8<sup>+</sup> T cells than did the spleens of mice vaccinated with KP cells expressing only mLAMA4 (Fig. 4d). The differences in efficacy between mixed cellular vaccines and dual antigen-expressing KP.mLAMA4.mITGB1 vaccines support previous findings that effective vaccines are those in which the MHC-I and MHC-II epitopes reside on the same peptide strand, potentially leading to more efficient uptake and presentation of both antigens by the same antigen-presenting cell (APC)<sup>20,35</sup>. A similar situation would be expected to occur when both antigens were present in the same tumour cell used for vaccination.

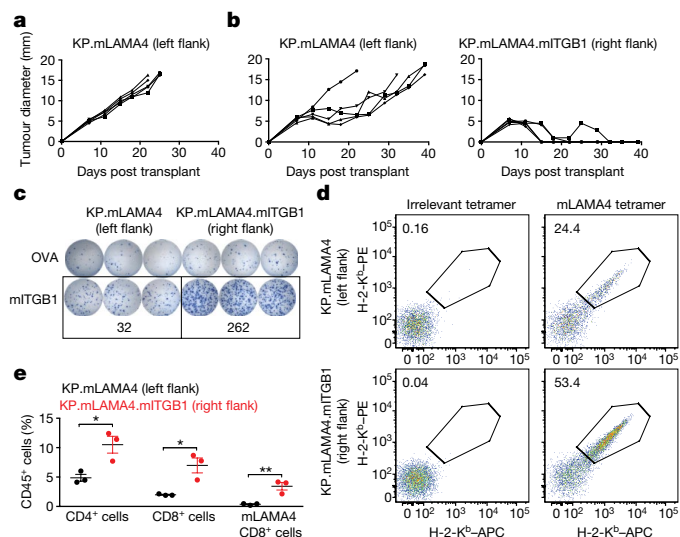
## MHC-II antigen expression at tumour site

To investigate whether CD4<sup>+</sup> T cells are required beyond the priming and maturation of anti-tumour CTLs, we tested whether tumour cell expression of MHC-II neoantigens was necessary at the site of tumour rejection. We assessed the in vivo growth of contralaterally injected KP.mLAMA4.mITGB1 and KP.mLAMA4 tumours in either immunodeficient or immunocompetent mice treated with ICT. The contralateral tumours grew at equivalent rates in *Rag2*<sup>-/-</sup> mice (Extended Data Fig. 9a). However, ICT treatment of wild-type mice bearing contralateral tumours resulted in complete rejection of the KP.mLAMA4.mITGB1 tumour but only delayed outgrowth of the KP.mLAMA4 tumour on



**Fig. 4 | MHC-II neoantigens are required for optimal tumour vaccine efficacy.** **a**, Schematic of tumour vaccine strategy. Naive syngeneic 129S6 mice were vaccinated with  $5 \times 10^5$  lethally irradiated KP sarcoma cells expressing the indicated antigens. Ten days after vaccination, mice were injected with  $2 \times 10^6$  T3 sarcoma cells on the opposite flank and growth or rejection of T3 tumours was monitored. **b**, Growth curves of T3 sarcoma cells in vaccinated mice as in **a**. Data are individual tumour diameters from mice injected in three independent experiments (numbers in figure represent number rejected over total number  $n$  for each group). **c**, Kaplan–Meier curves showing survival of mice in **b**.  $P$  values calculated using Mantel–Cox test. **d**, ELISPOT analysis of 1  $\mu$ M peptide-pulsed splenocytes 10 days after vaccination of naive mice with irradiated KP.mLAMA4 or KP.mLAMA4.mITGB1 cells as in **a**. Data from three independent experiments are shown as mean  $\pm$  s.e.m. number of spots. \*\*\* $P=0.0002$  (unpaired, two-tailed  $t$ -test).





**Fig. 5 | Expression of an MHC-II neoantigen by tumour cells has localized impact on tumour composition.** **a**, Wild-type syngeneic 129S4 mice were injected with  $1 \times 10^6$  KP.mLAMA4 cells followed by treatment with anti-PD-1 + anti-CTLA4 on days 3, 6, and 9 after transplantation. Representative data from one of three individual experiments are shown as individual tumour diameters ( $n = 5$  per group per experiment). **b**, Mice were injected on opposite sides with  $1 \times 10^6$  KP.mLAMA4 cells and  $1 \times 10^6$  KP.mLAMA4.mITGB1 cells followed by treatment as in **a**. Representative data from one of three individual experiments are shown as individual tumour diameters ( $n = 5$  per group per experiment). **c**, Mice were injected as in **b** and IFN $\gamma$  ELISPOT analysis of tumour-infiltrating CD4 $^{+}$  T cells stimulated with naive splenocytes pulsed with  $2 \mu$ g ml $^{-1}$  of the indicated peptides was performed 11 days after transplantation. Numbers beneath images indicate the average number of spots in mITGB1-stimulated wells from three independent experiments. **d**, Tetramer staining of mLAMA4-specific CD8 $^{+}$  TILs 11 days after transplantation of mice in **b**. Representative data from one of four independent experiments are shown as per cent of mLAMA4-specific cells within the CD8 $^{+}$  T cell population. **e**, Quantification of tumour-infiltrating T cells from mice in **b** 11 days after transplantation. Data are shown as per cent of total viable CD45 $^{+}$  cells  $\pm$  s.e.m. \* $P = 0.02$ , \*\* $P = 0.009$  (unpaired, two-tailed  $t$ -test).

the opposite flank (Fig. 5a, b). This result shows that CTLs specific for mLAMA4 can control tumours expressing both the cognate MHC-I epitope and the helper MHC-II epitope locally, but function poorly against distant but related tumours that lack CD4 neoepitopes. In similar experiments, we investigated whether mITGB1-specific CD4 $^{+}$  T cells generated from KP.mLAMA4.mITGB1 tumours were sufficient to control the outgrowth of KP.mITGB1 tumours on the opposite flank. In this setting, contralateral KP.mITGB1 tumour growth was identical to that observed in mice bearing only a single KP.mITGB1 tumour (Extended Data Fig. 9b, c). Together, these results show that tumour cell expression of MHC-II-restricted neoantigens and the presence of tumour-specific CD4 $^{+}$  T cells in the tumour microenvironment are required to maintain tumour control during ICT but are not sufficient to mediate tumour rejection by themselves.

To expand this observation, we investigated whether CD4 $^{+}$  T cells and expression of MHC-II neoantigens in tumour cells are required to maintain functional CD8 $^{+}$  T cell memory. When mice that had been cured of T3 tumours by ICT treatment were rechallenged with T3 tumour cells, they rejected the cells. However, if mice were depleted of CD4 $^{+}$  T cells before being rechallenged, they did not control T3 tumour outgrowth (Extended Data Fig. 9d). In parallel experiments, mice previously cured of KP.mLAMA4.mITGB1 tumours by surgical resection were protected against subsequent rechallenge with KP.mLAMA4.mITGB1 but were unable to prevent outgrowth of KP.mLAMA4 or KP9025 tumours

(Extended Data Fig. 9e). Thus, both expression of MHC-II neoantigens by tumour cells and CD4 $^{+}$  T cell help are required for the maintenance of tumour-specific immunologic memory.

Last, we investigated whether an MHC-II tumour neoantigen can significantly affect the local tumour microenvironment (gating strategy, Extended Data Fig. 10a). The expression of inducible nitric oxide synthase (iNOS) is higher in macrophages that populate tumours destined to be rejected after ICT than in macrophages from progressively growing tumours, and this expression is induced by ICT-dependent production of IFN $\gamma$ <sup>33</sup>. iNOS $^{+}$  macrophages were present at threefold higher levels in ICT-treated KP.mLAMA4.mITGB1 tumours than in contralateral KP.mLAMA4 tumours (Extended Data Fig. 9g, h). ELISPOT analysis of tumour-infiltrating CD4 $^{+}$  T cells showed 5.9-fold more IFN $\gamma$  $^{+}$  mITGB1-specific CD4 $^{+}$  T cells in KP.mLAMA4.mITGB1 tumours than in contralateral KP.mLAMA4 tumours (Fig. 5c, Extended Data Fig. 9f). Flow cytometry analysis of the lymphoid compartment (gating strategy, Extended Data Fig. 10b) identified 3.7-fold more CD8 $^{+}$  T cells and 9-fold more mLAMA4-specific CD8 $^{+}$  T cells in KP.mLAMA4.mITGB1 tumours than in KP.mLAMA4 tumours (Fig. 5d, e). We then investigated whether CD4 $^{+}$  T cells were sufficient to mediate these changes, by comparing iNOS $^{+}$  macrophages in KP.mLAMA4.mITGB1 tumours with those in contralateral KP.mITGB1 tumours. KP.mLAMA4.mITGB1 tumours contained 83-fold more iNOS $^{+}$  macrophages than did KP.mITGB1 tumours (Extended Data Fig. 9i, j). Together, these data show that MHC-II-restricted anti-tumour responses are necessary but not sufficient in ICT-sensitive tumour models to induce localized effects on the immune composition of tumours.

## Discussion

The work described herein focuses on the functional role of MHC-II restricted tumour neoantigens in mediating ICT-dependent anti-tumour responses in a well-characterized mouse sarcoma model. Using an HMM-based tool (hmMHC), we have predicted and validated that an N710Y point mutation in the integrin ITGB1 forms a major MHC-II restricted neoepitope of the T3 MCA sarcoma. It is reasonable that mITGB1 represents a major MHC-II neoantigen of T3 tumour cells because ITGB1 is the second most highly expressed mutation in T3 tumour cells and the point mutation in mITGB1 generates a novel anchor residue that promotes high affinity binding to I-A $^b$ . Moreover, others have proposed that secreted tumour proteins are favoured targets for CD4 $^{+}$  T cell responses because they are more easily taken up by professional APCs<sup>36</sup>. Localization of mITGB1 on the cell membrane would also be likely to facilitate efficient access by APCs, although we did not directly test this idea. Notably, we do not rule out the possibility that T3 cells express other MHC-II-restricted epitopes that might be elicited by vaccination<sup>18,19</sup>. Nevertheless, we have shown that mITGB1 functions as a major neoantigen of T3 cells during naturally occurring anti-tumour responses.

By defining authentic MHC-I and MHC-II neoantigens of T3 sarcoma cells, we have shown that, in a minimal antigen system, a single clonally expressed MHC-I neoantigen (mLAMA4) and a single clonally expressed MHC-II neoantigen (mITGB1) are necessary and sufficient to render nonimmunogenic, oncogene-driven KP9025 sarcoma cells sensitive to ICT. Using KP9025 sarcoma cells that express different combinations of mLAMA4 and/or mITGB1, we have shown that CD4 $^{+}$  T cell responses are required for optimal priming of MHC-I restricted CD8 $^{+}$  T cells and their maturation into CTLs, in either the presence or absence of ICT. We have also shown that optimal anti-tumour responses occur when tumour cells express both MHC-I and MHC-II neoantigens. In part, this requirement reflects the potential need for CD4 $^{+}$  T cell responses in the tumour microenvironment and, from previous work, appears to be at least partially due to production of IFN $\gamma$  by tumour-specific CD4 $^{+}$  T cells<sup>33</sup>. We find it of particular interest that the generation of effective tumour immunity requires MHC-II neoantigens following either vaccination with tumour-specific neoantigen vaccines or ICT. These results provide new insights into the role of MHC-II neoantigens in natural and therapeutic immune

responses to tumours. They also suggest that patients with tumours that are predicted to contain immunogenic MHC-I neoantigens or have favourable tumour mutational burdens could still be unresponsive to immunotherapies, owing to the absence of immunogenic MHC-II-restricted CD4<sup>+</sup> T cell antigens. This possibility has not been critically evaluated yet, owing to the past absence of reliable MHC-II prediction algorithms. Future work is needed to test this hypothesis in patients with cancer undergoing immunotherapy.

**Note added in proof:** As this Article was being prepared for publication, an independent paper was published online describing an MHC-II prediction algorithm for human tumours<sup>37</sup>.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1671-8>.

- Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science* **331**, 1565–1570 (2011).
- Larkin, J. et al. Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *N. Engl. J. Med.* **373**, 23–34 (2015).
- Motzer, R. J. et al. Nivolumab versus everolimus in advanced renal-cell carcinoma. *N. Engl. J. Med.* **373**, 1803–1813 (2015).
- Borghaei, H. et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *N. Engl. J. Med.* **373**, 1627–1639 (2015).
- Lennerz, V. et al. The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. *Proc. Natl Acad. Sci. USA* **102**, 16013–16018 (2005).
- Matsushita, H. et al. Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature* **482**, 400–404 (2012).
- DuPage, M., Mazumdar, C., Schmidt, L. M., Cheung, A. F. & Jacks, T. Expression of tumour-specific antigens underlies cancer immunoediting. *Nature* **482**, 405–409 (2012).
- Robbins, P. F. et al. Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* **19**, 747–752 (2013).
- Gubin, M. M. et al. Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* **515**, 577–581 (2014).
- Wölfel, T. et al. A p16INK4a-insensitive CDK4 mutant targeted by cytolytic T lymphocytes in a human melanoma. *Science* **269**, 1281–1284 (1995).
- Snyder, A. et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).
- Strønen, E. et al. Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. *Science* **352**, 1337–1341 (2016).
- Rizvi, N. A. et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
- Van Allen, E. M. et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207–211 (2015).
- Spranger, S. et al. Density of immunogenic antigens does not explain the presence or absence of the T-cell-inflamed tumor microenvironment in melanoma. *Proc. Natl Acad. Sci. USA* **113**, E7759–E7768 (2016).
- Hugo, W. et al. Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* **165**, 35–44 (2016).
- Hellmann, M. D. et al. Genomic features of response to combination immunotherapy in patients with advanced non-small-cell lung cancer. *Cancer Cell* **33**, 843–852 (2018).
- Kreiter, S. et al. Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* **520**, 692–696 (2015).
- Ott, P. A. et al. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).
- Ossendorp, F., Mengedé, E., Camps, M., Filius, R. & Melief, C. J. Specific T helper cell requirement for optimal induction of cytotoxic T lymphocytes against major histocompatibility complex class II negative tumors. *J. Exp. Med.* **187**, 693–702 (1998).
- Corthay, A. et al. Primary antitumor immune response mediated by CD4<sup>+</sup> T cells. *Immunity* **22**, 371–383 (2005).
- Wong, S. B. J., Bos, R. & Sherman, L. A. Tumor-specific CD4<sup>+</sup> T cells render the tumor environment permissive for infiltration by low-avidity CD8<sup>+</sup> T cells. *J. Immunol.* **180**, 3122–3131 (2008).
- Bos, R. & Sherman, L. A. CD4<sup>+</sup> T-cell help in the tumor milieu is required for recruitment and cytolytic function of CD8<sup>+</sup> T lymphocytes. *Cancer Res.* **70**, 8368–8377 (2010).
- Zhu, Z. et al. CD4<sup>+</sup> T cell help selectively enhances high-avidity tumor antigen-specific CD8<sup>+</sup> T cells. *J. Immunol.* **195**, 3482–3489 (2015).
- Borst, J., Ahrends, T., Bąbata, N., Melief, C. J. M. & Kastenmüller, W. CD4<sup>+</sup> T cell help in cancer immunology and immunotherapy. *Nat. Rev. Immunol.* **18**, 635–647 (2018).
- Andreatta, M. et al. An automated benchmarking platform for MHC class II binding prediction methods. *Bioinformatics* **34**, 1522–1528 (2018).
- Mittal, P. et al. Tumor-unrelated CD4 T cell help augments CD134 plus CD137 dual costimulation tumor therapy. *J. Immunol.* **195**, 5816–5826 (2015).
- Suri, A., Walters, J. J., Rohrs, H. W., Gross, M. L. & Unanue, E. R. First signature of islet  $\beta$ -cell-derived naturally processed peptides selected by diabetogenic class II MHC molecules. *J. Immunol.* **180**, 3849–3856 (2008).
- Tran, E. et al. Cancer immunotherapy based on mutation-specific CD4<sup>+</sup> T cells in a patient with epithelial cancer. *Science* **344**, 641–645 (2014).
- Linnemann, C. et al. High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4<sup>+</sup> T cells in human melanoma. *Nat. Med.* **21**, 81–85 (2015).
- Old, L. J. & Boyse, E. A. Immunology of experimental tumors. *Annu. Rev. Med.* **15**, 167–186 (1964).
- Wei, S. C. et al. Distinct cellular mechanisms underlie anti-CTLA-4 and anti-PD-1 checkpoint blockade. *Cell* **170**, 1120–1133 (2017).
- Gubin, M. M. et al. High-dimensional analysis delineates myeloid and lymphoid compartment remodeling during successful immune-checkpoint cancer therapy. *Cell* **175**, 1014–1030 (2018).
- Marzo, A. L. et al. Tumor-specific CD4<sup>+</sup> T cells have a major “post-licensing” role in CTL mediated anti-tumor immunity. *J. Immunol.* **165**, 6047–6055 (2000).
- Bennett, S. R. M., Carbone, F. R., Karamalis, F., Miller, J. F. A. P. & Heath, W. R. Induction of a CD8<sup>+</sup> cytotoxic T lymphocyte response by cross-priming requires cognate CD4<sup>+</sup> T cell help. *J. Exp. Med.* **186**, 65–70 (1997).
- Corthay, A., Lundin, K. U., Lørvik, K. B., Hofgaard, P. O. & Bogen, B. Secretion of tumor-specific antigen by myeloma cells is required for cancer immunosurveillance by CD4<sup>+</sup> T cells. *Cancer Res.* **69**, 5901–5907 (2009).
- Abelin, J. G. et al. Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. *Immunity* <https://doi.org/10.1016/j.immuni.2019.08.012> (2019).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### Mice

Male wild-type 129S6 mice (for experiments involving T3 cells) were purchased from Taconic Farms. Male wild-type 129S4 mice (for experiments involving KP9025 cells) and 129S6 *Rag2*<sup>-/-</sup> mice were bred in our specific-pathogen free facility. All in vivo experiments were performed in our specific-pathogen free facility and used mice between the ages of 8 and 12 weeks. All experiments were performed in accordance with procedures approved by the AAALAC-accredited Animal Studies Committee of Washington University in St Louis and were in compliance with all relevant ethical regulations.

### Tumour transplantation

T3 MCA-induced sarcoma cells were previously generated in our laboratory in 129S6 wild-type mice. KP sarcoma cell lines were provided by T.J., and were generated following intramuscular injection of lentiviral Cre-recombinase into 129S4 *Kras*<sup>LSL-G12D/+</sup> × *Tp53*<sup>fl/fl</sup> mice. Tumour cells were cultured in Roswell Park Memorial Institute (RPMI) medium (Hyclone) supplemented with 10% fetal calf serum (FCS) (Hyclone). Cell lines were authenticated using whole-exome sequencing and verification of specific antigen expression. All cell lines used tested negative for mycoplasma contamination. For transplantation, cells were washed extensively in PBS and resuspended at a density of  $13.34 \times 10^6$  cells per ml (T3) or  $6.67 \times 10^6$  cells per ml (KP sarcomas) in PBS. Then, 150 µl was injected subcutaneously into the rear flanks of syngeneic recipient mice. For irradiated tumour cell vaccines, KP.mLAMA4, KP.mITGB1 or KP.mLAMA4.mITGB1 sarcoma cells were lethally irradiated with 10 Gy and 500,000 cells were injected subcutaneously into 129S6 mice. T3 challenge following vaccination occurred on the opposite flank. Following tumour transplantation, animals were randomly assigned to treatment groups. No statistical methods were used to determine group size. Tumour growth was measured by calipers and individual growth curves are represented as the average of two perpendicular diameters. Tumour measurements were performed blinded to treatment group. In accordance with our IACUC-approved protocol, maximal tumour diameter was 20 mm in one direction, and in no experiments was this limit exceeded.

### Tumour rechallenge

For tumour rechallenge following surgical resection, primary tumours were allowed to grow until they reached 10 mm in size or to the time point indicated. Following surgical removal of the established tumour, animals were rested for 30 days. Animals were then rechallenged on the opposite flank with either the same tumour line as was used in the primary tumour challenge or the tumour line indicated. For tumour rechallenge following ICT-mediated rejection, primary tumours were rejected following treatment with combination anti-PD-1 and anti-CTLA4 ICT. After tumours were no longer apparent, animals were rested for 30 days followed by rechallenge on the opposite flank with the same tumour line as was used in the primary challenge or the tumour line indicated.

### Epitope prediction

The identification of point mutations in T3 and KP sarcomas and the prediction of MHC-I epitopes in KP and F244 sarcomas were performed as previously described<sup>9</sup>. To predict neoepitopes, we applied hmMHC, our newly developed HMM-based binding predictor, trained on the most recent IEDB data. HMMs inherently accommodate inputs of variable length and have already demonstrated reasonable performance for prediction of MHC binding affinity<sup>38</sup>. Our predictor uses a fully connected HMM with emissions representing amino acids (see a pedagogical example in Extended Data Fig. 1a). We trained the model on a set of known binders using the Baum–Welch algorithm<sup>39</sup>, as implemented by the GHMM library (<http://ghmm.sourceforge.net/>). A trained HMM

returns the likelihood of a peptide to be a binder, which we represent as the  $-\log_{10}$  odds ratio, where a smaller value indicates that a peptide has a higher likelihood of being a binder. The model that we apply in this study was trained on murine H2-I-A<sup>b</sup> binders taken from the IEDB full MHC ligand export (downloaded on 25 November 2018, containing 1,072,460 entries). Non-binders were not used in model training. The categorization of the data into binders and non-binders was based on the qualitative and quantitative fields of IEDB entries: binders are peptides with  $IC_{50} \leq 500$  nM or with positive, positive-high or positive-intermediate binding quality. These data came largely from mass spectrometry assays. We validated the model using the Monte Carlo (shuffle-split) cross-validation approach, with ten random partitions of H2-I-A<sup>b</sup> binders from IEDB into training and validation sets, with a relative validation set size of 0.2. As the number of non-binders in the IEDB dataset was insufficient for validation, we used decoy sets composed of random natural peptides as non-binders. Protein-coding transcript translation sequences for *Mus musculus* were obtained from GENCODE release M19 (GENCODE project, 2018); there are 65,257 translations. For every cross-validation partition, the translations were randomly cut into fragments uniformly distributed in the interval [12, 24], which generated about  $1.5 \times 10^6$  fragments. Of this set of random natural peptides, a random sample 100 times the number of binders in the validation set was taken. The 100-fold bias in the number of generated non-binders and uniform distribution of their lengths are in line with recent work on MHC binding prediction, in particular netMHCpan-4.0<sup>40</sup>. We have also performed experiments in which the distribution of random natural peptide lengths followed the distribution of lengths in the IEDB dataset (Extended Data Fig. 1d) and found no significant difference in results in our setting compared to uniform distribution. The rationale for the 100-fold bias is that for a sample of peptide fragments from an organism, it is commonly considered that about 1–2% will bind to MHC receptors. On average, there were 4,412 binders in a training set, and 771 binders and 77,086 random natural peptides in a validation set. Classification performance of our predictor was significantly higher than the performance of the two best-known class II binding predictors<sup>41</sup> (netMHCII-2.3 and netMHCIIpan-3.2), compared on our ten validation datasets. This is due, in part, to the large amount of new mass spectrometry data compared to the data on which the recent netMHCII(pan) predictors were trained (netMHCIIpan-3.2 public dataset available at <http://www.cbs.dtu.dk/suppl/immunology/NetMHCIIpan-3.2/> contains 1,794 measurements for H-2-I-A<sup>b</sup>, all qualitative, of which 431 are binders and 1,363 are weak or non-binders). We do not exclude the possibility that netMHCII(pan), as a method, performs better than the HMM method. As the published netMHCII(pan) tools lack re-training capability, we cannot compare the methods and draw conclusions on netMHCII(pan) performance on new qualitative data. We determined the threshold for strong binders by calibrating the predictor to return a percentile rank against a large decoy set of random natural peptides. We used the approach taken by the existing neural network-based predictors, in which strong binders are predictions in the second percentile of the empirical distribution of predictions on random natural peptides<sup>40</sup>. The decoy set was generated from the mouse proteome in the same way as for validation and consists of about  $1.5 \times 10^6$  fragments with lengths in the interval [12, 24]. Predicted neoantigens were further prioritized using the NER: the ratio between the binding predictions for the mutant and wild-type peptides. Expression of each mutation is represented as FPKM generated from cDNA capture sequencing.

### Peptides

All 27-mer peptides used for neoantigen screening (Supplementary Table 1) were purchased from Peptide 2.0 and purified by high-performance liquid chromatography (HPLC) to >95% purity. The T3-specific mutant amino acid was placed in the centre of the peptide and was flanked on both sides with 13 amino acids of wild-type peptide sequence.

## ELISPOT

Cells from tumours or lymph nodes were enriched for CD4<sup>+</sup> or CD8<sup>+</sup> T cells using the Miltenyi mouse CD4<sup>+</sup> or CD8<sup>+</sup> enrichment kits according to the manufacturer's protocols. Ten thousand TIL-derived T cells or 50,000 tumour-draining lymph node (TDLN)-derived T cells were stimulated with 500,000 splenocytes isolated from naive mice pulsed with 2 µg ml<sup>-1</sup> 29-mer peptide (class II) or 1 µM 15-mer peptide (class I). For analysis from spleens, 500,000 cells from whole-spleen preparations were used. Cells were stimulated overnight in anti-mouse IFN $\gamma$ -coated ELISPOT plates (Immunospot). Plates were developed according to the manufacturer's protocol and spots were quantified using a CTL ImmunoSpot S6 Universal machine and Professional 6.0.0 software.

## Mass spectrometry

For isolation of I-A<sup>b</sup>-bound peptides, 5 × 10<sup>8</sup> T3.CIITA cells were washed twice with PBS and snap-frozen. MHC-II molecules were isolated by immunoaffinity purification using the I-A<sup>b</sup>-specific antibody Y-3P (BioX-Cell) coupled to cyanogen bromide-activated sepharose 4B (GE Healthcare) as described<sup>42</sup>. Peptides were eluted with 0.2% trifluoroacetic acid, cleaned by detergent removal (Pierce Detergent Removal Spin Columns, Thermo Scientific) and desalting (Pierce C-18 Spin Columns, Thermo Scientific), dried, and resuspended in 2% acetonitrile (ACN) and 0.1% formic acid (20 µl). For mass spectrometry, a Dionex UltiMate 1000 system (Thermo Scientific) was coupled to an Orbitrap Fusion Lumos (Thermo Scientific) through an Easy-Spray ion source (Thermo Scientific). Peptide samples were loaded (15 µl/min, 3 min) onto a trap column (100 µm × 2 cm, 5 µm Acclaim PepMap 100 C18, 50 °C), eluted (200 nl/min) onto an Easy-Spray PepMap RSLC C18 column (2 µm, 50 cm × 75 µm ID, 50 °C, Thermo Scientific) and separated with the following gradient (all percentages indicate buffer B: 0.1% formic acid in ACN): 0–110 min, 2–22%; 110–120 min, 22–35%; 120–130 min, 35–95%; 130–150 min, isocratic at 95%; 150–151 min, 95–2%; 151–171 min, isocratic at 2%. Spray voltage was 1,900 V, ion transfer tube temperature was 275 °C, and RF lens was 30%. Mass spectrometry scans were acquired in profile mode (375–1,500 Da at 120,000 resolution (at *m/z* 200)); centroided HCD MS/MS spectra were acquired using a Top Speed method (charge states 2–7, 3 s cycle time, threshold 2 × 10<sup>4</sup>, quadrupole isolation (0.7 Da), 30,000 resolution, collision energy 30%) with dynamic exclusion enabled (5 ppm, 60 s). Raw data files were uploaded to PEAKS X (Bioinformatics Solutions) for processing, de novo sequencing and database searching against the UniProtKB/Swiss-Prot Mouse Proteome database (downloaded 1 December 2019; 22,286 entries), appended with a truncated sequence of mITGB1 (±20 amino acids from the site of mutation), with mass error tolerances of 10 ppm and 0.01 Da for parent and fragment, respectively, no enzyme specificity, and methionine oxidation as a variable modification. False discovery rate (FDR) estimation was enabled, and proteins were filtered for  $-\log_{10} P \geq 0$  and one unique peptide to give 1% FDR at the peptide-spectrum match level. Peptides matching to mITGB1 were manually verified by visual inspection.

## Antibodies

For immune checkpoint therapy, rat IgG2a anti-PD1 (RMP1-14, Leinco) and mouse IgG2b anti-CTLA4 (9D9, Leinco Technologies) antibodies were used. Mice were injected intraperitoneally with 200 µg of each antibody on days 3, 6 and 9 after tumour transplantation. For multi-colour flow cytometry, we used antibodies against CD45 (30-F11), CD11B (M1/70), THY1.2 (30H12), CD4 (RM4-5), CD8 $\beta$  (YTS156.7.7), I-E/I-A (M5/114.15.2), CD64 (X54-5/7.1), LY6G (1A8), T-BET (4B10), CD150/SLAM (TC15-12F12.2), KLRG1 (2F1), ICOS (15F9), CD44 (IM7), PD-1 (29F.1A12), SIINFEKL-H-2-K<sup>b</sup> (25-D1.16) (BioLegend), CD24 (M1/69), F4/80 (T45-2342) (BD Biosciences), FOXP3 (FJK-16 s, eBiosciences) and iNOS (CXNFT, Invitrogen). Zombie NIR (BioLegend) was used to stain for cellular viability. The BD Cytotfix/Cytoperm Plus kit (BD Biosciences) was used according

to the manufacturer's protocol for intracellular staining of iNOS, T-BET and FOXP3.

## Tetramer staining

Tetramer staining for mLAMA4-specific CD8<sup>+</sup> T cells was performed as previously described<sup>9</sup>. I-A<sup>b</sup> monomers bound to CLIP or mITGB1 were a gift from K. Wucherpfennig. For staining, biotinylated pI-A<sup>b</sup> monomers were labelled at a 4:1 molar ratio with streptavidin-APC or streptavidin-PE (Prozyme). One million cells from whole-tumour digests were stained with equal amounts of APC and PE tetramer at 20 µg ml<sup>-1</sup> for 2 h at room temperature. Tetramer staining was stabilized through the use of anti-PE and anti-APC cells beads (Miltenyi), similar to previously published methods for MHC-I tetramers<sup>43</sup>, followed by surface staining for CD11B, THY1.2 and CD4.

## Multi-cytokine assay

CD4<sup>+</sup> T cells were enriched from tumours 12 days after transplantation using the Miltenyi mouse CD4<sup>+</sup> enrichment kit. Ten thousand enriched CD4<sup>+</sup> T cells were stimulated in serum-free medium with 500,000 splenocytes isolated from naive mice pulsed with 2 µg ml<sup>-1</sup> peptide. Following a 24-h incubation, secretion of IL-10, IL-1B, IL-2, IL-4, IL-5, IL-6, IL-22, IL-9, IL-13, IL-27, IL-23, IFN $\gamma$ , IL-12 p70, GM-CSF, TNF, IL-17A and IL-18 was measured using a flow-based ProcartaPlex T<sub>H</sub>1/T<sub>H</sub>2/T<sub>H</sub>9/T<sub>H</sub>17/T<sub>H</sub>22/T<sub>reg</sub> cytokine panel (Luminex Technologies) following the manufacturer's protocol.

## Plasmids

Full-length mLAMA4 and mITGB1 were cloned from T3 cDNA and full-length CIITA was cloned from 129S6 splenocytes. Gene blocks encoding SIINFEKL and the minimal epitope of mSB2 were purchased from Integrated DNA Technologies. All constructs were cloned into the BglIII site of pMSCV-IRES GFP (mLAMA4, CIITA, and mSB2) or pMSCV (mITGB1 and SIINFEKL) using the Gibson Assembly method (New England Biolabs). To generate neoantigen-expressing KP sarcoma cell lines and T3.CIITA cells, constructs were transiently transfected into Phoenix Eco cells using Eugene (Promega). After 48 h, viral supernatants were subsequently used for transfection of KP sarcoma line 9025 or T3 cells. KP.mLAMA4, KP.mITGB1, KP.mLAMA4.mITGB1, KP.mSB2.SIINFEKL and T3.CIITA clones were obtained by limiting dilution.

## CD4<sup>+</sup> T cell hybridomas and CTLL assay

Bulk CD4<sup>+</sup> T cells from T3 tumours were isolated 12 days after transplantation and stimulated with lethally irradiated T3.CIITA cells to establish a rapidly dividing cell line. CD4<sup>+</sup> T cells were fused with BW5147 cells and cloned via limiting dilution. To assess antigen specificity and to map the mITGB1 MHC-II binding core, splenocytes were collected from naive mice and pulsed with 10 µg ml<sup>-1</sup> peptide unless otherwise stated. Fifty thousand hybridoma cells were incubated with 100,000 peptide-pulsed splenocytes overnight and culture medium was collected. Production of IL-2 was assayed by proliferation-dependent thymidine incorporation using the IL-2 dependent CTLL-2 cell line. Data are represented as counts per million (cpm).

## Measuring IFN $\gamma$ production by CD8<sup>+</sup> T cell clones

Tumour cells were treated with 100 U ml<sup>-1</sup> IFN $\gamma$  for 48 h before use. One hundred thousand CTL cells specific against mLAMA4 (74.17) or mSB2 (C3) were co-cultured with 50,000 tumour cells for 48 h. IFN $\gamma$  in supernatants was quantified using an IFN $\gamma$  ELISA kit (eBioscience) according to the manufacturer's protocol.

## In vivo cytotoxicity assay

For targets, splenocytes were collected from naive mice, stained with either 5 µM or 0.5 µM CFSE (CFSE<sup>hi</sup> and CFSE<sup>lo</sup>) (Thermo Fisher Scientific) and pulsed with either mLAMA4 (CFSE<sup>hi</sup>) or SIINFEKL (CFSE<sup>lo</sup>) peptide, respectively, at 1 µM overnight. Cells were washed extensively and

# Article

combined at a 50:50 ratio in PBS, and  $20 \times 10^6$  cells were injected retro-orbitally into tumour-bearing mice 11 days after tumour transplantation. Naive, non-tumour bearing mice were used as a control. Spleens from tumour-bearing or control naive mice were removed 24 h after cell transfer, stained with Zombie NIR viability dye (Biolegend) and quantified for the presence of CFSE-labelled target cells. On histograms, equivalent heights of CFSE<sup>hi</sup> and CFSE<sup>lo</sup> peaks indicate that equivalent numbers of each cell population are present, and that no cytotoxicity was observed. Peaks that differ in height, where the CFSE<sup>lo</sup> population is more abundant than the CFSE<sup>hi</sup> population, indicate that cytotoxicity was observed specifically against the mLAMA4 peptide-pulsed, CFSE<sup>hi</sup> population of cells. The equation used for calculating per cent specific lysis was  $[1 - (\text{naive control ratio/experimental ratio})] \times 100$  with ratio = irrelevant percentage/specific epitope percentage.

## Statistics

Statistical analysis was performed using GraphPad Prism software version 7. Unless otherwise noted, significance was determined with an unpaired, two-tailed Student's *t*-test.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

Nucleotide variant calls generated from cDNA capture sequencing of the T3 and KP9025 sarcoma lines and used in the prediction of antigens shown in Fig. 1a, Extended Data Fig. 3a, b and 6b are provided as Supplementary Data 1 and Supplementary Data 2.

## Code availability

Code for the hmMHC algorithm used to predict presentation of neoantigens by I-A<sup>b</sup> can be accessed at <https://github.com/artiomovlab/hmmhc>.

38. Mamitsuka, H. Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins* **33**, 460–474 (1998).
39. Welch, L. R. Hidden Markov models and the Baum–Welch algorithm. *IEEE Inf. Theory Soc. Newsl.* **53**, 10–13 (2003).
40. Jurtz, V. et al. NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* **199**, 3360–3368 (2017).
41. Jensen, K. K. et al. Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* **154**, 394–406 (2018).
42. Kowalewski, D. J. & Stevanović, S. Biochemical large-scale identification of MHC class I ligands. *Methods Mol. Biol.* **960**, 145–157 (2013).
43. Tungatt, K. et al. Antibody stabilization of peptide–MHC multimers reveals functional T cells bearing extremely low-affinity TCRs. *J. Immunol.* **194**, 463–474 (2015).

**Acknowledgements** We thank all members of the Schreiber laboratory for discussions and technical support. This work was supported by grants to R.D.S. from the National Cancer

Institute of the National Institutes of Health (R01CA190700), the Parker Institute for Cancer Immunotherapy, the Cancer Research Institute, Janssen Pharmaceutical Company of Johnson and Johnson and the Prostate Cancer Foundation, and by a Stand Up to Cancer–Lustgarten Foundation Pancreatic Cancer Foundation Convergence Dream Team Translational Research Grant. Stand Up to Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. E.A. and D.M.L. were supported by a postdoctoral training grant (T32 CA00954729) from the National Cancer Institute. D.M.L. and M.M.G. were supported by the Irvington Postdoctoral Fellowship from the Cancer Research Institute. M.D. is a St Baldrick's Scholar with support from Hope with Hazel and a Pew–Stewart Scholar for Cancer Research supported by the Pew Charitable Trusts. J.P.W. is supported by the National Cancer Institute of the National Institutes of Health Paul Calabresi Career Development Award in Clinical Oncology (K12CA167540). M.M.G. is supported by a Parker Bridge Scholar Award from the Parker Institute for Cancer Immunotherapy. K.W.W. receives support from the National Institutes of Health (R01CA238039). T.J. receives support from a National Institutes of Health Cancer Center Support Grant (P30CA14051) and the Howard Hughes Medical Institute. E.R.U. receives support from the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (A114551 and DK058177). Aspects of the studies, including ELISPOT, were performed by D. Bender at the Immunomonitoring Laboratory (IML), which is supported by the Andrew M. and Jane N. Bursky Center for Human Immunology and Immunotherapy Programs and the Alvin J. Siteman Comprehensive Cancer Center which, in turn, is supported by a National Cancer Institute of the National Institutes of Health Cancer Center Support Grant (P30CA91842).

**Author contributions** E.A. conceived and designed the experiments, collected the data, performed and interpreted the analyses, and wrote the manuscript. D.M.L. and A.P.M. planned experiments, and collected and analysed data. I.K. conceived of and designed the hmMHC algorithm and performed analyses using it, and wrote the methodological description. M.D. generated the KP9025 sarcoma cell line. A.M.L. provided technical assistance and helped to plan experiments using MHC-II tetramers. W.M. and C.F.L. planned, performed and analysed mass spectrometry experiments. E.E. assisted with bioinformatics analyses. A.N.V. assisted with the generation of the CD4<sup>+</sup> T cell hybridomas, and helped to design and perform experiments using them. D.R. designed experiments involving multi-colour flow cytometry and collected and analysed the data. J.P.W. provided technical support for MHC-I tetramer staining. M.M.G. assisted in experiment planning. R.F.V.M. collected and analysed data for experiments involving multi-colour flow cytometry. C.D.A., K.C.F.S. and J.M.W. provided technical assistance throughout the study. A.C. collected data. K.W.W. provided mITGB1–MHC-II monomers and provided assistance in experimental design. T.J. provided support in experimental design and data analysis regarding the KP9025 sarcoma line. M.N.A. conceived and designed the hmMHC algorithm and provided bioinformatics support. E.R.U. provided assistance with experimental design. R.D.S. conceived experiments, interpreted data, and wrote the manuscript. All authors contributed to manuscript revision.

**Competing interests** R.D.S. is a cofounder, scientific advisory board member, stockholder, and royalty recipient of Jounce Therapeutics and Neon Therapeutics and is a scientific advisory board member for A2 Biotherapeutics, BioLegend, Codiak Biosciences, Constellation Pharmaceuticals, NGM Biopharmaceuticals and Sensei Biotherapeutics. K.W.W. serves on the scientific advisory board of Tscan Therapeutics and Nextechinvest, and receives sponsored research funding from Bristol–Myers Squibb and Novartis; these activities are not related to the findings described in this publication. T.J. is a member of the Board of Directors of Amgen and Thermo Fisher Scientific. He is also a co-founder of Dragonfly Therapeutics and T2 Biosystems, and serves on the Scientific Advisory Board of Dragonfly Therapeutics, SQZ Biotech, and Skyhawk Therapeutics. None of these affiliations represent a conflict of interest with respect to the design or execution of this study or interpretation of data presented in this manuscript. The laboratory of T.J. currently also receives funding from the Johnson & Johnson Lung Cancer Initiative and Calico, but this funding did not support the research described in this manuscript.

## Additional information

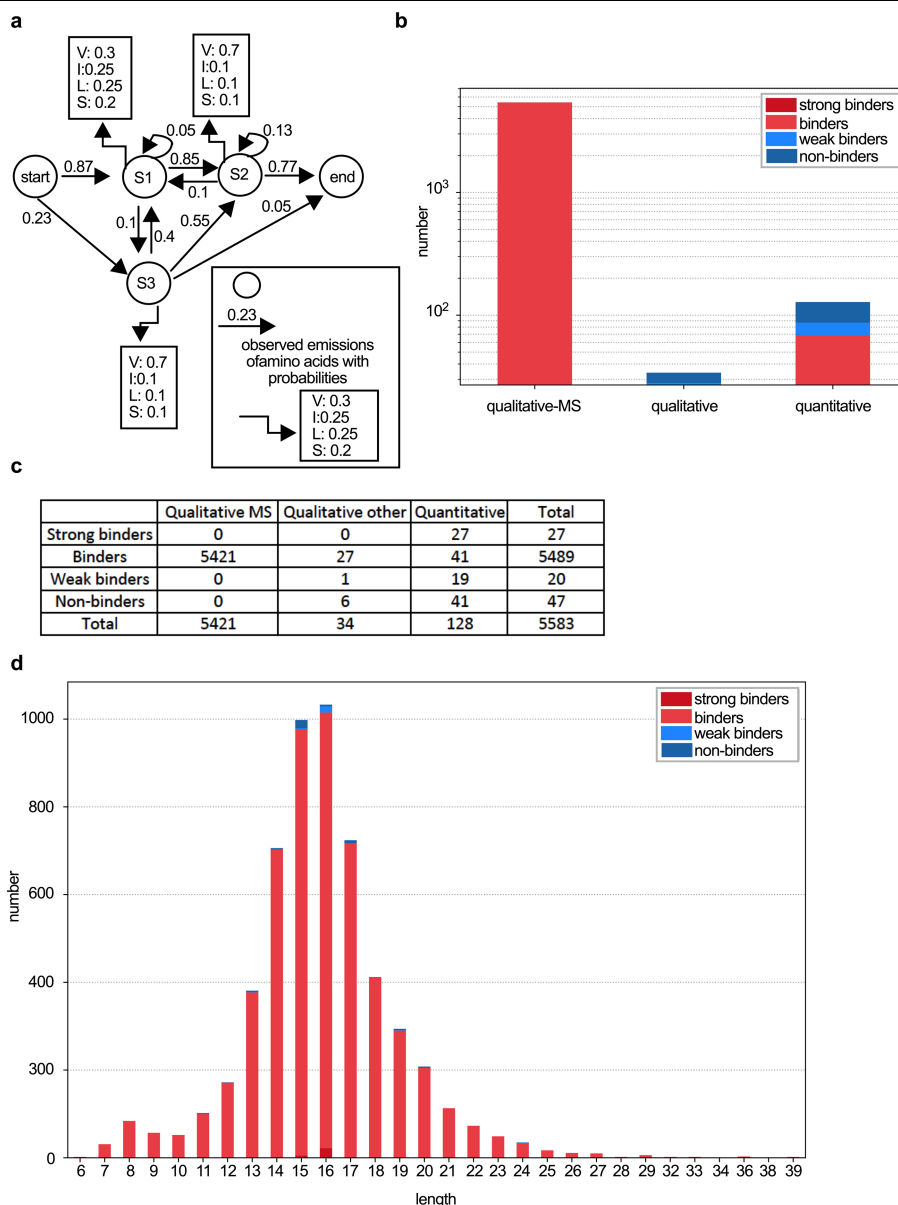
**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1671-8>.

**Correspondence and requests for materials** should be addressed to R.D.S.

**Peer review information** Nature thanks Lelia Delamarre, Cornelis Melief and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

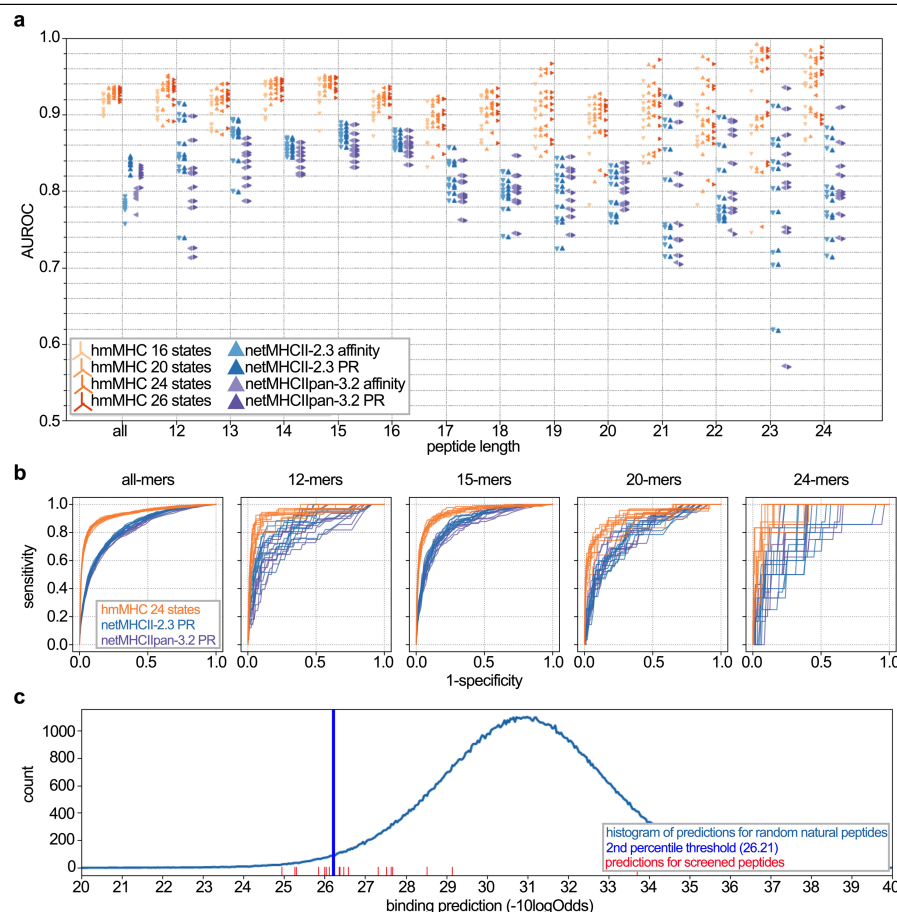
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.





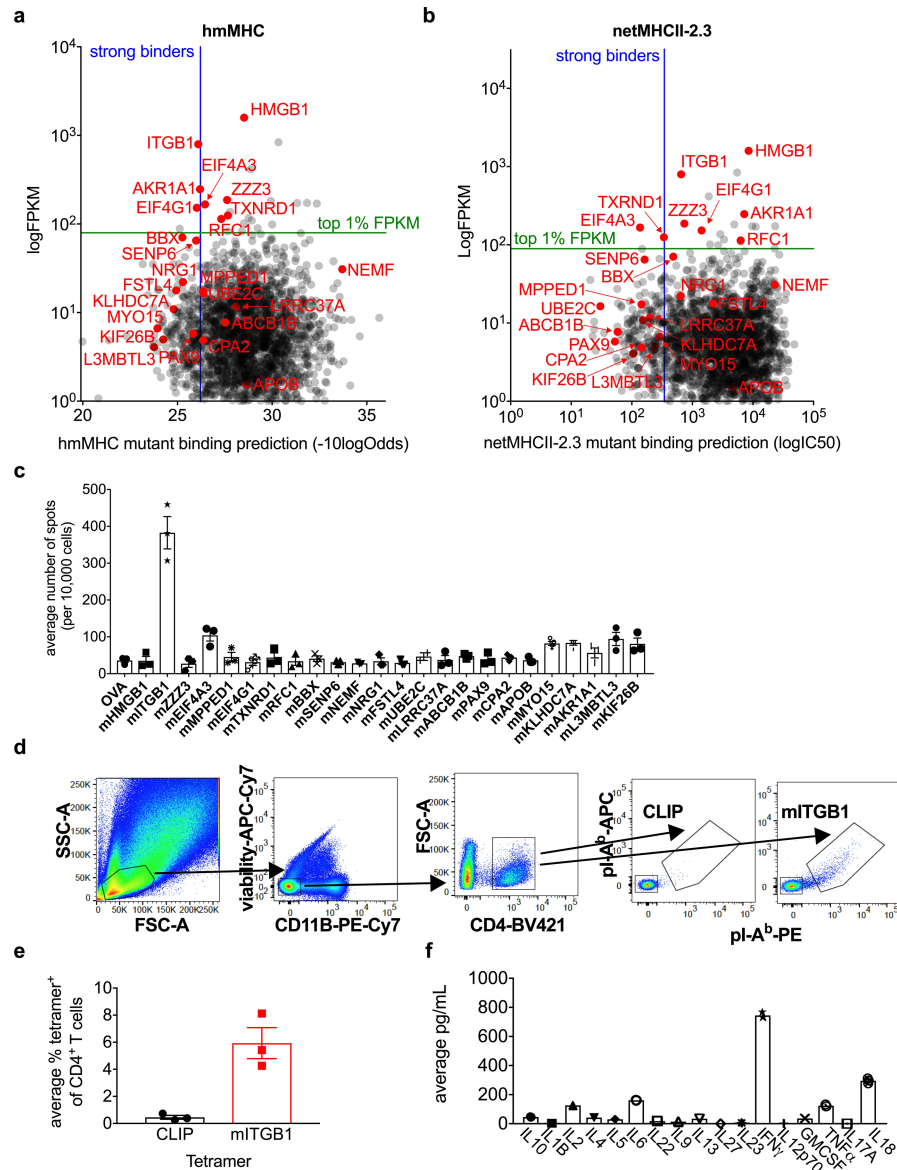
**Extended Data Fig. 1 | The hmMHC predictive algorithm and IEDB'18 H2-I-A<sup>b</sup> training dataset composition.** **a**, An example of a fully connected HMM with three hidden states, and emissions corresponding to amino acids. **b-d**, Composition of IEDB dataset (MHC full ligand export downloaded on 25

November 2018) represented as number of peptides per binding category and measurement type (**b, c**) and binding category and peptide length (**d**). Strong binders:  $IC_{50} \leq 50$  nM; binders:  $50 \text{ nM} < IC_{50} \leq 500$  nM; weak binders:  $500 \text{ nM} < IC_{50} \leq 5,000$  nM; non-binders: all remaining peptides.



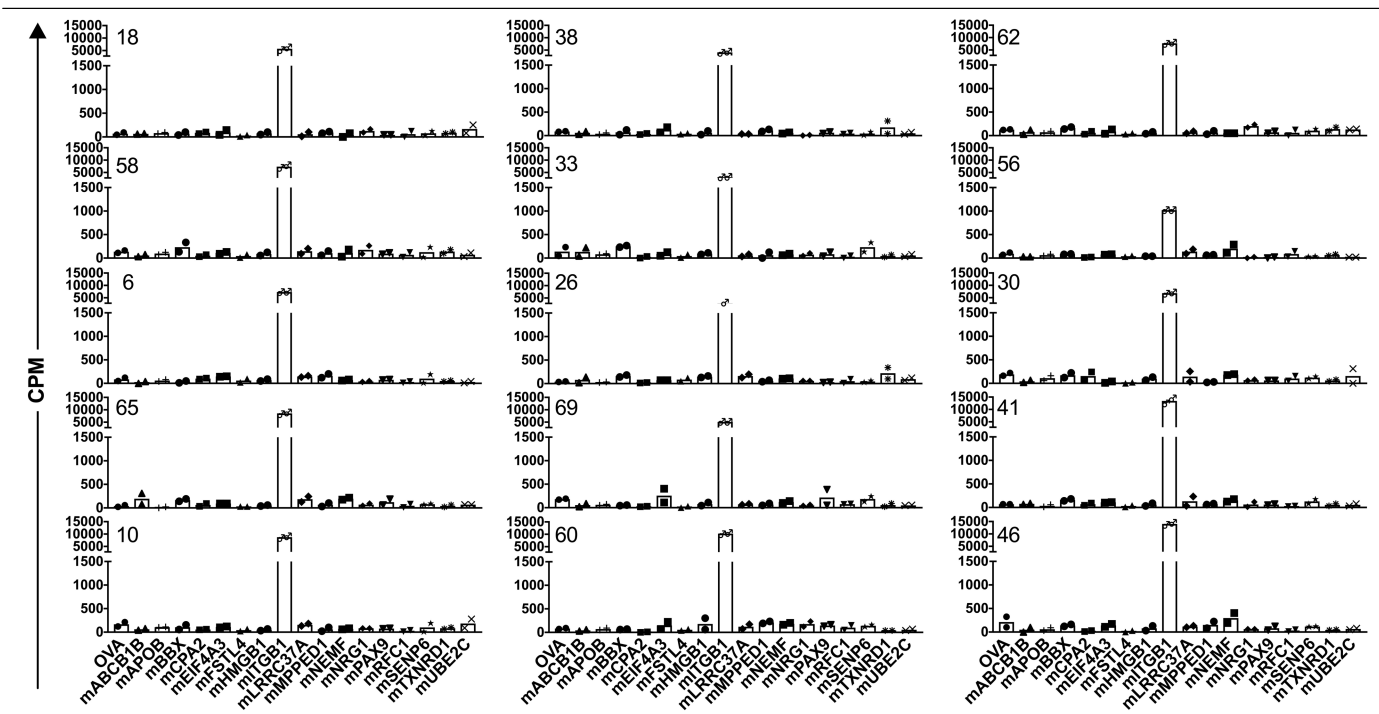
**Extended Data Fig. 2 | Performance of hmMHC compared to netMHCII-2.3 and netMHCIIpan-3.2.** **a**, hmMHC (orange shapes) underwent 10× cross-validation. In each of the ten cross-validation partitions, on average there were 4,412 binders in the training set, and 771 binders and 77,086 random natural peptides in the validation set. Performance was compared in terms of AUROC to the performance of netMHCII-2.3 (blue triangles) and netMHCIIpan-3.2 (purple triangles) applied on the same validation sets. For hmMHC, performance for different numbers of hidden states is shown. For netMHCII-2.3 and

netMHCIIpan-3.2, performance is shown for both predicted affinity and percentile rank (PR). **b**, Receiver operating characteristic (ROC) curves showing the performance of hmMHC on the H2-I-A<sup>b</sup> dataset compared to existing predictors. ROC curves of all peptides and per specific peptide length for every cross-validation partition are shown. **c**, Illustration of percentile rank for strong binder classification calibrated on random natural peptides. Red lines indicate the percentile ranks of peptides screened for CD4<sup>+</sup> T cell reactivity.



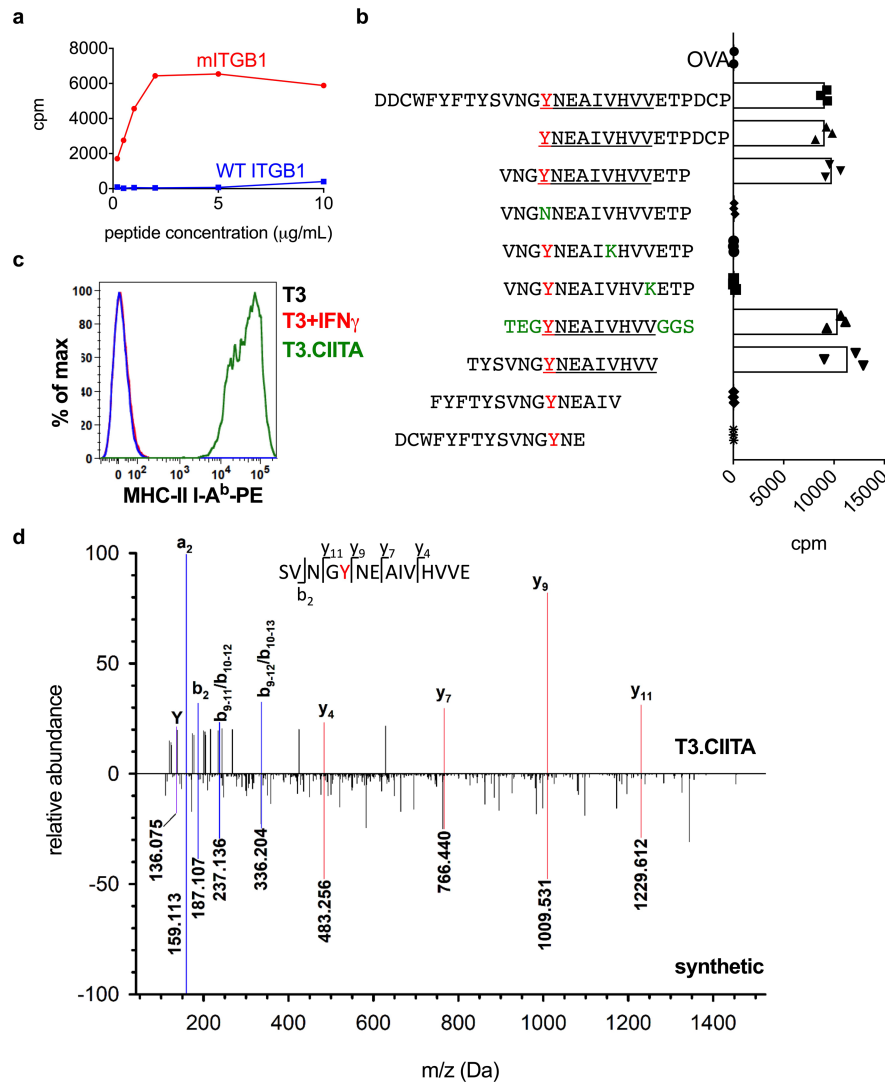
**Extended Data Fig. 3 | mITGB1 is a major MHC-II-restricted neoantigen in T3 sarcomas.** **a, b**, T3 MHC-II neoantigen predictions for all expressed mutations were made using hmMHC (**a**) and netMHCII-2.3 (**b**) (netMHCIIpan-3.2 predictions yielded very similar results, data not shown). The predictions are shown as  $-\log_{10}$  odds predictor value or  $\log IC_{50}$  (smaller values indicate higher likelihood of being presented by I-A<sup>b</sup>) and expression level (FPKM). Strong binders are defined as mutations residing in the second percentile of I-A<sup>b</sup> binding predictions for random natural peptides for each algorithm ( $-\log_{10}$  odds  $\leq 26.21$  or  $\log IC_{50} \leq 343.8$  nM). The N710Y mutation in ITGB1 met the strong binder threshold in the hmMHC predictions but not in the netMHCII-2.3 predictions. Red dots indicate all mutations that were screened for CD4<sup>+</sup> T cell reactivity. Green line denotes high-expression cut-off (FPKM = 89.1). Blue line indicates strong binder cut off for each algorithm. **c**, Two million T3 sarcoma cells were injected subcutaneously into syngeneic mice and CD4<sup>+</sup> TILs were

isolated on day 12. IFN $\gamma$  ELISPOT was performed using naive splenocytes pulsed with  $2 \mu\text{g ml}^{-1}$  of the indicated peptides. Data are shown as mean of three independent experiments  $\pm$  s.e.m. **d**, Gating strategy for pI-A<sup>b</sup> tetramer staining of whole TILs. **e**, Quantification of mITGB1-tetramer and CLIP-tetramer staining of CD4<sup>+</sup> T cells from whole T3 TILs 12 days after transplantation. Data are shown as mean  $\pm$  s.e.m. per cent tetramer-positive cells of CD4<sup>+</sup> cells from three independent experiments. **f**, Syngeneic 129S6 mice were injected subcutaneously with  $2 \times 10^6$  T3 sarcoma cells and TIL-derived CD4<sup>+</sup> T cells were collected 12 days after transplantation. CD4<sup>+</sup> T cells were stimulated with naive splenocytes pulsed with  $2 \mu\text{g ml}^{-1}$  OVA<sub>323-339</sub> control or mITGB1<sub>697-724</sub> peptide for a flow-based multi-cytokine array. Representative data from one of two independent experiments using pools of five tumours each are shown as average of technical triplicate wells from three pooled tumours.



**Extended Data Fig. 4 | T3 TIL-derived CD4<sup>+</sup> T cell hybridomas are reactive against mLRRC37.** CTL assay of T3 TIL-derived CD4<sup>+</sup> T cell hybridoma lines stimulated with naive splenocytes pulsed with 2  $\mu\text{g ml}^{-1}$  of the individual

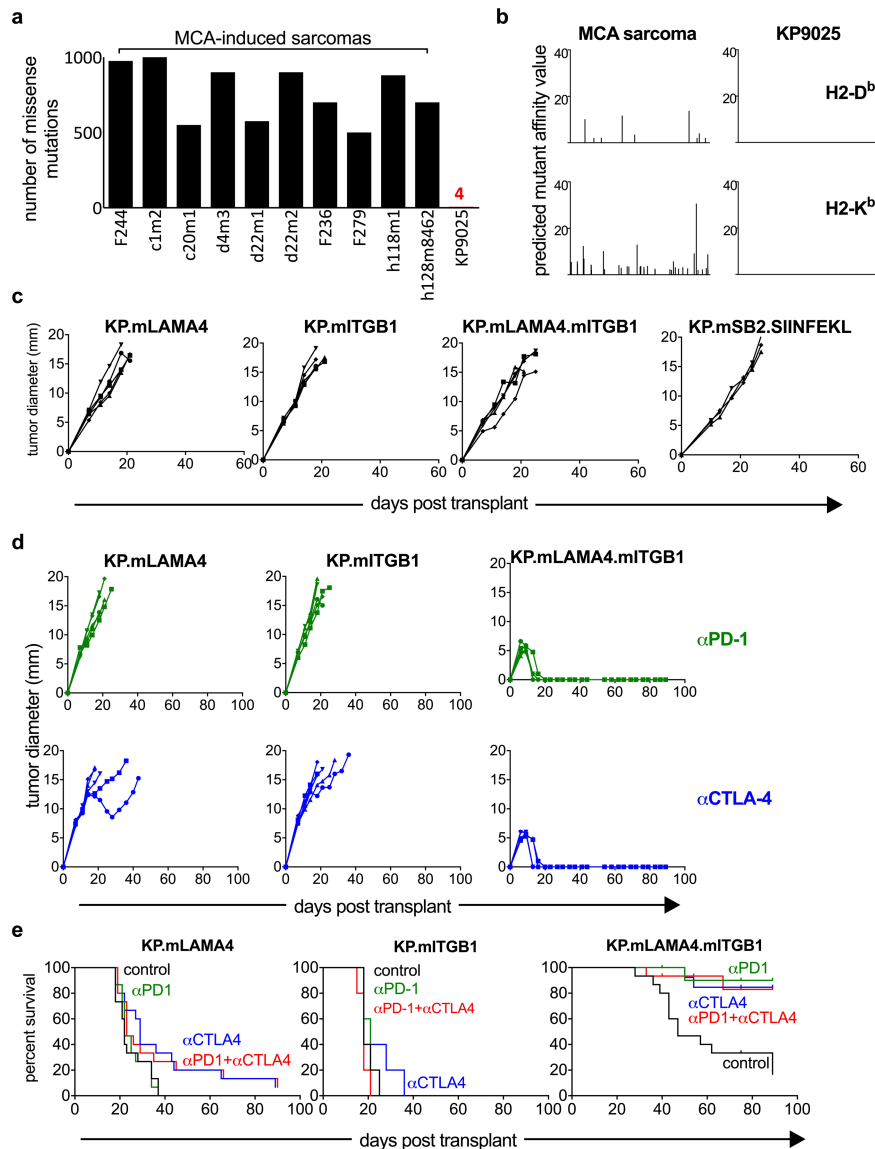
indicated peptides. Representative data from one of three independent experiments are shown as average cpm from technical duplicate wells.



**Extended Data Fig. 5 | The mITGB1 epitope is presented on I-A<sup>b</sup>.** **a**, T3 CD4<sup>+</sup> T cell hybridomas were stimulated with  $2 \mu\text{g mL}^{-1}$  mITGB1(710Y) or wild-type ITGB1(710N) peptide-pulsed splenocytes. Activation was measured by CTLL assay. Representative data from three independent hybridoma lines are shown as average of technical replicate wells. **b**, Mapping of the mITGB1 MHC-II binding core was performed using the CD4<sup>+</sup> T cell hybridoma line 41 stimulated with naive splenocytes pulsed with  $2 \mu\text{g mL}^{-1}$  of overlapping peptides covering mITGB1<sub>697-724</sub>. Red denotes the T3-specific mutant amino acid at position p1 of the minimal epitope; underlining denotes the validated binding core. Green amino acids represent random residue substitutions used to specifically define valines at residues 715 and 718 as the p6 and p9 MHC-II binding positions and the

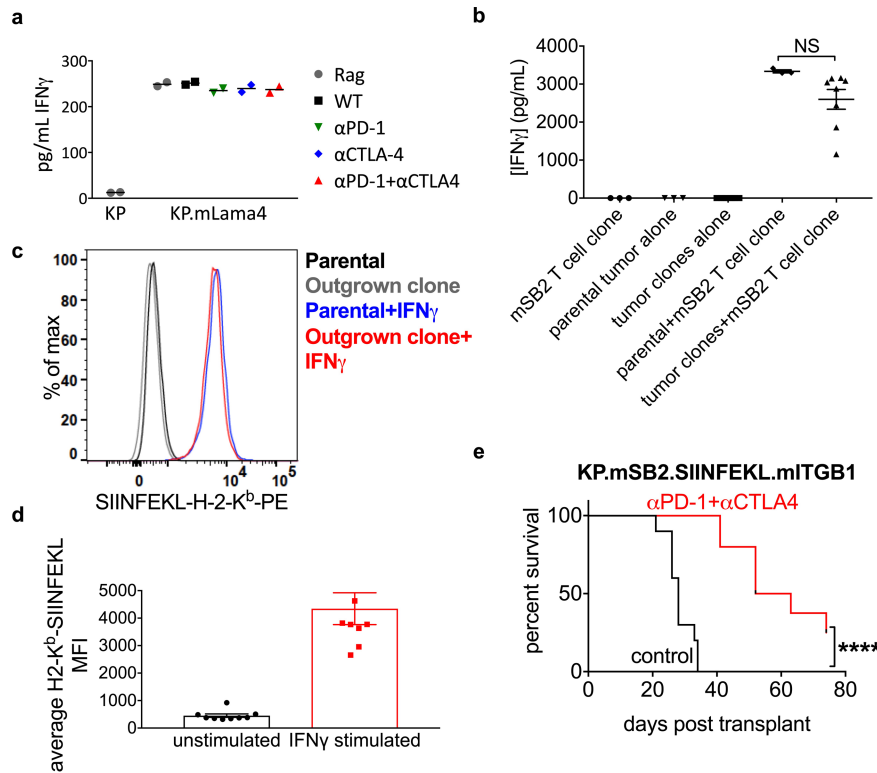
complete MHC-II binding core. Representative data from two independent experiments are shown as the average of technical triplicate wells. **c**, MHC-II I-A<sup>b</sup> staining of parental T3 cells, IFN $\gamma$ -stimulated T3 cells and T3 cells transduced with a vector encoding CIITA (T3.CIITA). Representative data from one of three independent experiments are shown. **d**, Mirror plot showing match between MS/MS spectra of the 14-mer peptide sequence encompassing the N710Y site of mITGB1 eluted from T3.CIITA cells (positive axis) and a corresponding synthetic peptide (negative axis). Labelled m/z values reflect those experimentally observed for the endogenous peptide, with peaks representing b ions highlighted in blue and y ions in red.





**Extended Data Fig. 6 | mITGB1 CD4<sup>+</sup> T cells are required for tumour rejection in response to ICT. a**, Comparison of total number of expressed missense mutations between ten different MCA-induced sarcomas and KP9025 cells. Mutations were defined by whole-exome sequencing and RNA sequencing, and mutational load is shown on a per cell basis. **b**, Comparison of predicted neoantigen MHC-I affinity values between KP9025 and MCA-induced sarcoma F244 for H-2D<sup>b</sup> (top) and H-2K<sup>b</sup> (bottom). KP9025 cells were not predicted to express any MHC-I neoantigens. **c**, *Rag2*<sup>-/-</sup> mice were subcutaneously injected with  $1 \times 10^6$  KP.mLAMA4, KP.mITGB1, KP.mLAMA4.mITGB1 or KP.mSB2.SIINFEKL cells. Representative data from one of two independent experiments are

presented as tumour diameters from individual mice ( $n = 5$  mice per group for KP.mLAMA4, KP.mITGB1 and KP.mLAMA4.mITGB1 and  $n = 3$  mice for KP.mSB2.SIINFEKL group per experiment). **d**, Wild-type syngeneic 129S4 mice were injected subcutaneously with  $1 \times 10^6$  KP.mLAMA4, KP.mITGB1 or KP.mLAMA4.mITGB1 cells and treated with anti-PD-1 (top) or anti-CTLA single agent ICT (bottom) on days 3, 6, and 9 after transplantation. Representative data from one of three independent experiments are shown as tumour diameters from individual mice ( $n = 5$  in all groups per experiment). **e**, Survival curves for experiments in **d** and Fig. 2e ( $n = 15$  in all groups).



**Extended Data Fig. 7 | Outgrowth of nonimmunogenic sarcoma cells expressing MHC-I neoantigens is not a result of cancer immunoediting.**

**a**, *Rag2*<sup>+/+</sup> or wild-type 129S4 mice were injected with  $1 \times 10^6$  KP9025 or KP.mLAMA4 cells and treated with anti-PD-1, anti-CTLA or anti-PD-1 + anti-CTLA4 on days 3, 6 and 9 after injection. Tumours were removed once they reached a maximum diameter of 20 mm in any direction and sarcoma cell lines were established ex vivo. Cell lines were stimulated with IFN $\gamma$  to upregulate MHC-I and subsequently used to stimulate the mLAMA4-specific CD8<sup>+</sup> 74.14 T cell clone. Secretion of IFN $\gamma$  by T cells was measured using enzyme-linked immunosorbent assay (ELISA). Representative data from two independent experiments are represented as the average of two independent tumour samples in each group.

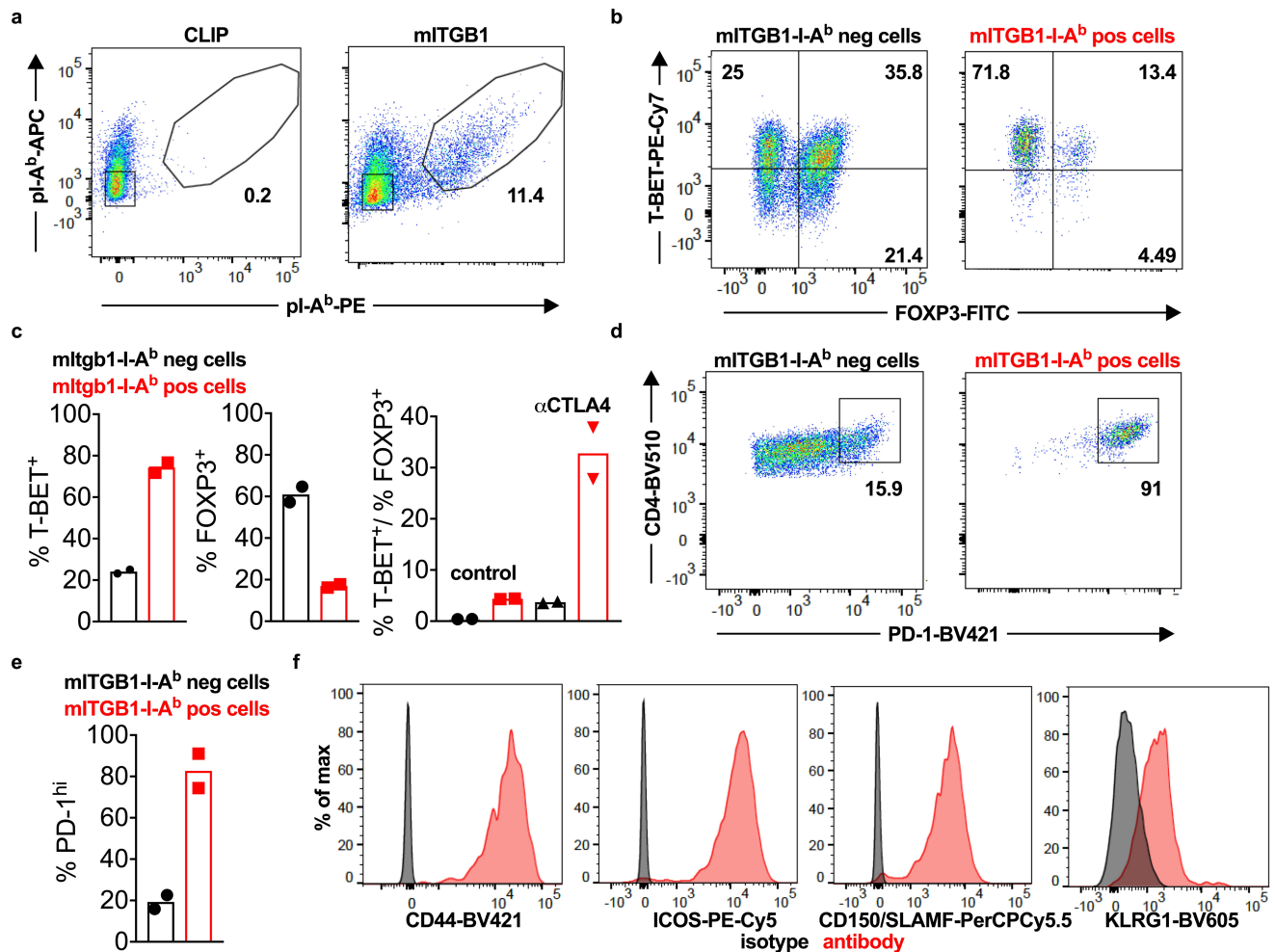
**b**, Wild-type 129S4 mice were injected with  $1 \times 10^6$  KP.mSB2.SIINFEKL cells and treated with anti-PD-1 + anti-CTLA4 combination ICT on days 3, 6 and 9 after injection. Tumours were removed as in **a**. Established ex vivo cell lines were

cloned by limiting dilution and parental KP.mSB2.SIINFEKL cells or individual clones from outgrown tumours were used to stimulate the mSB2-specific C3 CD8<sup>+</sup> T cell clone; production of IFN $\gamma$  was quantified using ELISA. Representative data from four independent experiments are presented as average IFN $\gamma$  concentration of eight individual clones  $\pm$  s.e.m. Significance was determined using an unpaired, two-tailed *t*-test.

**c**, Cell surface staining of SIINFEKL-H-2-K<sup>b</sup> expressed by unstimulated or IFN $\gamma$ -stimulated parental KP.mSB2.SIINFEKL cells or individual clones described in **b**. A representative histogram is shown.

**d**, Quantification of mean  $\pm$  s.e.m. SIINFEKL-H-2-K<sup>b</sup> mean fluorescence intensity (MFI) from eight individual clones in **c**. NS, not significant.

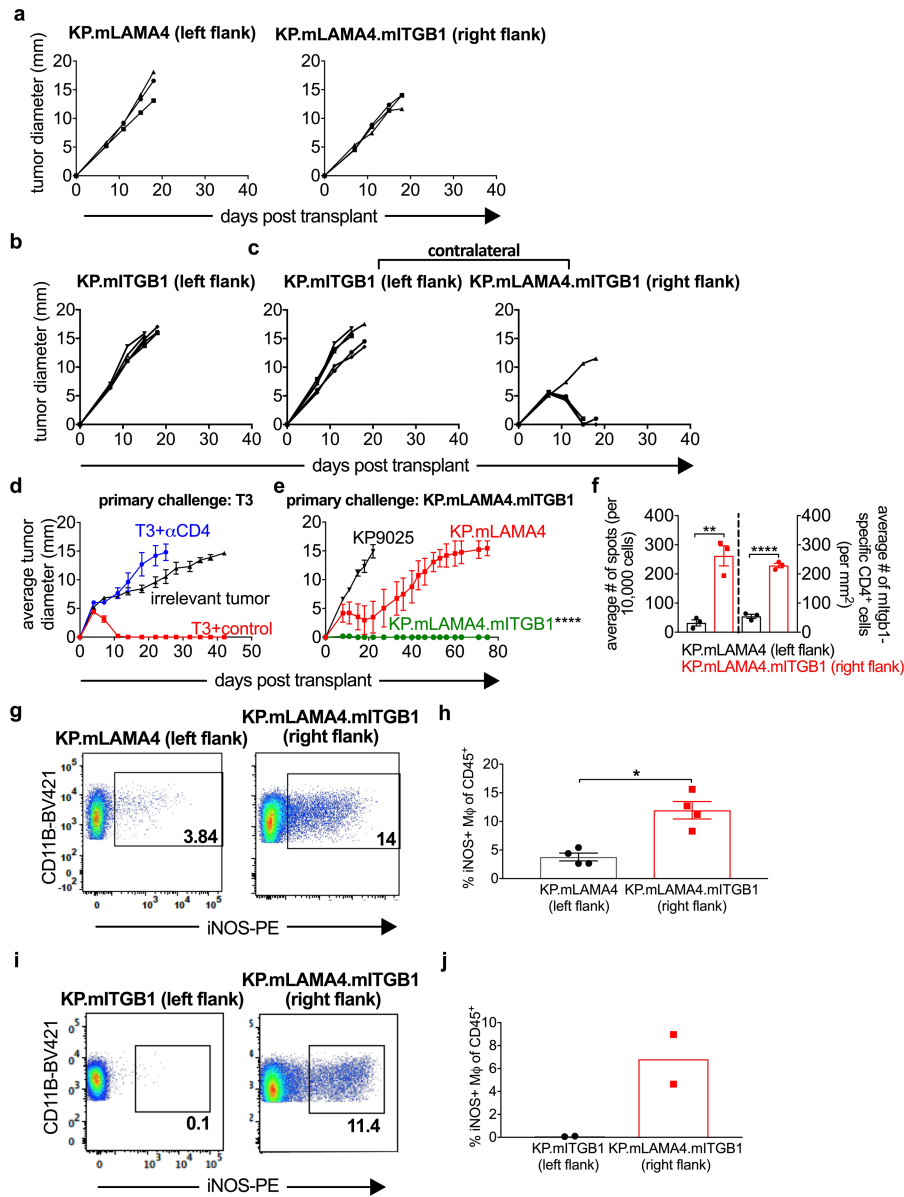
**e**, Survival curves of wild-type 129S4 mice injected subcutaneously with  $1 \times 10^6$  KP.mSB2.SIINFEKL.mITGB1 cells. Mice were treated with control monoclonal antibodies or anti-PD-1 + anti-CTLA4 combination ICT on days 3, 6 and 9 after injection. *n* = 10 mice per group from two independent experiments. \*\*\*\**P* =  $1.5 \times 10^{-5}$ , Mantel-Cox test.



**Extended Data Fig. 8 | mITGB1-specific CD4<sup>+</sup> T cells display an activated T<sub>H</sub>1 phenotype.** **a**, Whole TILs from KP.mLAMA4.mITGB1 tumours 12 days after transplantation were stained with mITGB1-I-A<sup>b</sup> tetramers. Populations were previously gated on viable CD11b<sup>+</sup> CD4<sup>+</sup> cells. Representative data from one of two independent experiments with five pooled tumours each are shown. **b**, mITGB1-I-A<sup>b</sup> tetramer-negative and tetramer-positive cells described in **a** were analysed for expression of T-BET and FOXP3. Representative plots are shown. **c**, Quantification of two independent experiments in **b** as average per cent of tetramer-negative and tetramer-positive cells staining positive for the indicated

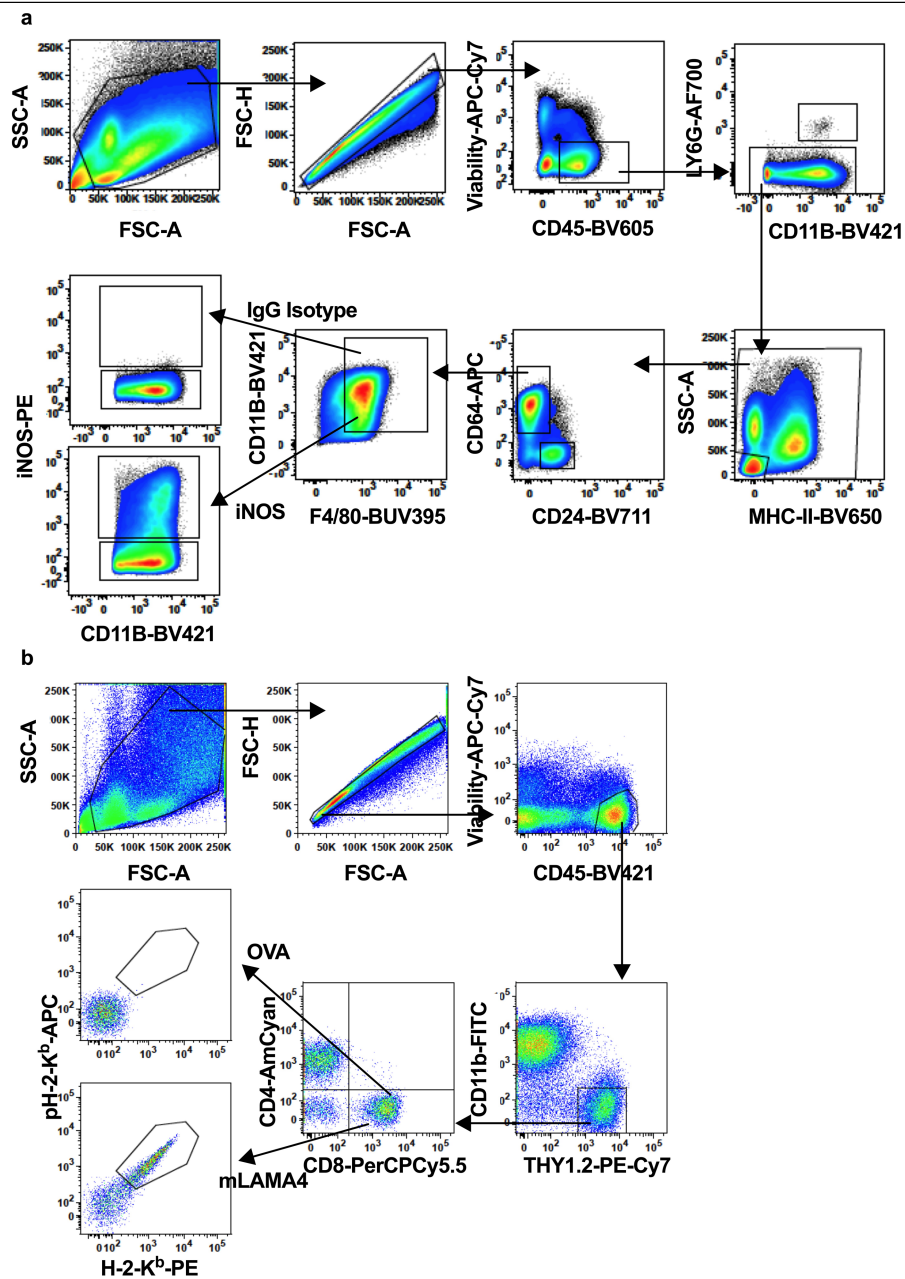
protein. Tumour-bearing mice were treated with control monoclonal antibodies or anti-CTLA4 on days 3, 6, and 9 after transplantation where indicated.

**d**, mITGB1-I-A<sup>b</sup> tetramer-positive and tetramer-negative cells in **a** were analysed for expression of PD-1. Representative plots are shown. **e**, Quantification of two independent experiments described in **d** shown as average per cent of tetramer-negative and tetramer-positive cells staining positive for PD-1. **f**, mITGB1-I-A<sup>b</sup> tetramer-positive cells in **a** were analysed for expression of the indicated proteins. Representative histograms from one of two independent experiments using pools of five tumours each are shown.



**Extended Data Fig. 9 | CD4<sup>+</sup> T cell help is required at the tumour site during primary and memory responses.** **a**, *Rag2*<sup>-/-</sup> mice were simultaneously injected with  $1 \times 10^6$  KP.mLAMA4 and KP.mLAMA4.mITGB1 cells on opposite flanks. Representative data from one of two independent experiments are shown as individual tumour diameter ( $n = 3$  in each experiment). **b**, Wild-type 129S4 mice were injected with  $1 \times 10^6$  KP.mITGB1 cells and were treated with anti-PD-1 + anti-CTLA4 combination ICT on days 3, 6, and 9 after injection. Representative data from one of two individual experiments are shown as individual tumour diameters ( $n = 5$  in all experiments). **c**, Wild-type 129S4 mice were simultaneously injected with  $1 \times 10^6$  KP.mLAMA4 and KP.mLAMA4.mITGB1 cells on opposite flanks and treated as in **b**. Representative data from one of two individual experiments are shown as individual tumour diameters ( $n = 5$  in all experiments). **d**, Wild-type 129S6 mice were injected subcutaneously with  $2 \times 10^6$  T3 sarcoma cells and were treated with anti-PD-1 + anti-CTLA4 combination ICT on days 3, 6, and 9 after injection. Following tumour rejection and a 30-day recovery period, tumour-experienced mice were rechallenged with  $2 \times 10^6$  T3 cells in the presence of control monoclonal antibody or CD4-depleting antibody, or with irrelevant sarcoma cells. Representative data from one of two independent experiments are shown as average tumour diameter  $\pm$  s.e.m. ( $n = 5$  in all groups per experiment). **e**, Wild-type 129S4 mice were injected subcutaneously with  $1 \times 10^6$  KP.mLAMA4.mITGB1 cells followed by surgical

resection 10 days after transplantation. After a 30-day recovery period, tumour-experienced mice were rechallenged with  $1 \times 10^6$  KP9025, KP.mLAMA4.mITGB1, or KP.mLAMA4 cells. Representative data from one of two independent experiments are shown as average tumour diameter  $\pm$  s.e.m. ( $n = 5$  in all groups per experiment). \*\*\*\* $P = 2 \times 10^{-6}$  by two-way ANOVA with multiple comparisons and Bonferroni correction. **f**, Quantification of data from three independent experiments in Fig. 5c is shown as average number of spots  $\pm$  s.e.m. (left) and average number of mITGB1-specific CD4<sup>+</sup> cells  $\pm$  s.e.m. (right). \*\* $P = 0.003$ , \*\*\*\* $P = 7.2 \times 10^{-5}$  (unpaired, two-tailed *t*-test). **g**, CD45<sup>+</sup>Ly6G<sup>+</sup>MHCII<sup>+</sup>CD64<sup>+</sup>CD25<sup>-</sup>CD11b<sup>+</sup>F4/80<sup>+</sup> macrophages in TILs from mice bearing the indicated contralateral tumours were analysed for expression of iNOS 11 days after tumour transplant. Representative data from four independent experiments are shown. **h**, Quantification of iNOS<sup>+</sup> macrophages from experiments in **f** as a per cent of total CD45<sup>+</sup> cells. Data are shown as average  $\pm$  s.e.m. of four independent experiments. \* $P = 0.03$  by unpaired, two-tailed *t*-test. **i**, CD45<sup>+</sup>Ly6G<sup>+</sup>MHCII<sup>+</sup>CD64<sup>+</sup>CD25<sup>-</sup>CD11b<sup>+</sup>F4/80<sup>+</sup> macrophages from the indicated contralateral tumours described were isolated 11 days after transplantation and analysed for expression of iNOS. Representative plots from two independent experiments are shown. **j**, Quantification of iNOS<sup>+</sup> macrophages from two independent experiments in **h** is shown as average per cent of total CD45<sup>+</sup> cells.



**Extended Data Fig. 10 | Gating strategies for multi-colour flow cytometry.** Gating strategies for multi-colour flow cytometry analysis of tumour-infiltrating macrophage (a) and T cell (b) populations.



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Flow cytometry data was collected using BD FACS Diva software version 8.0. ELISPOT data was collected using a CTL ImmunoSpot S6 Universal machine and Professional 6.0.0 software. The code for the hmMHC algorithm can be accessed at <https://github.com/artiomovlab/hmmhc>.

Data analysis

Flow cytometry data was analyzed using FlowJo software version 10.2. Mass spectrometry data was analyzed using PEAKS X. All statistical analysis was performed using GraphPad Prism software version 7.0c.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Nucleotide variant calls generated from cDNA capture sequencing of the T3 and KP9025 sarcoma lines and used in the prediction of antigens shown in Figure 1a, Extended Data Figure 3a-b, and Extended Data Figure 6b are available within the article as Supplementary Data 1 and Supplementary Data 2.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined based on 25 years experience with these types of tumors in general, and 6 years with these specific cell lines.
Data exclusions	No data was excluded from analysis.
Replication	All experiments used in this study were replicated in at least 2 independent experiments; the majority of experiments were performed at least 3 times. All attempts at replication were successful.
Randomization	In each experiment, all mice were obtained from the same cohort, and injected with the same preparation of tumor cells. Cages of animals were then randomly assigned into treatment groups.
Blinding	In these experiments, one person injected tumors, but longitudinal measurements and antibody treatments were performed by three independent people.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	Antibodies used for multi-color flow cytometry were CD45 (30-F11), CD11b (M1/70), Thy1.2 (30H12), CD4 (RM4-5), CD8 $\beta$ (YTS156.7.7), I-E/I-A (M5/114.15.2), CD64 (X54-5/7.1), Ly6G (1A8), T-bet (4B10), CD150/SLAM (TC15-12F12.2), KLRG1 (2F1), ICOS (15F9), CD44 (IM7), PD-1 (29F.1A12), SIINFEKL-H-2-Kb (25-D1.16) (Biolegend), CD24 (M1/69), F4/80 (T45-2342) (BD Biosciences), Foxp3 (FJK-16s, eBiosciences) and iNOS (CXNFT, Invitrogen).
Validation	All antibodies used for flow cytometry are commercially available and validation materials are available on the appropriate websites. Tetramers were validated by staining cell populations from irrelevant tumors not expressing the indicated antigens.

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	The T3 sarcoma cell line was generated via subcutaneous injection of methylcholanthrene into 129S6 Rag2-deficient mice. Resulting tumors were used to establish a parental sarcoma cell line, which was further cloned by limiting dilution to produce T3 (Matsushita et al., Nature 2012; Gubin et al., Nature 2014). The KP9025 cell line was generated via intramuscular injection of lentiviral cre-recombinase into the flanks of 129S4 strain mice expressing the KRas G12D oncogene and a floxed p53 cassette. A resulting tumor was used to establish the KP9025 cell line.
Authentication	Cell lines were authenticated by whole exome sequencing and RNA-seq, and further validated by expression of particular antigens.
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination.

Commonly misidentified lines  
(See [ICLAC](#) register)

No commonly misidentified cell lines were used.

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

Experiments utilized both Rag2-deficient and wild type male mus musculus strains 129S6 and 129S4 between 8 and 10 weeks of age. It was necessary to match the sex and strain of mice to which the cell line was generated.

Wild animals

This study did not include wild animals.

Field-collected samples

This study did not include samples collected from the field.

Ethics oversight

All animal studies were performed with the approval of the Association for the Accreditation of Laboratory Animal Care-accredited Animal Studies Committee of Washington University in St. Louis.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Flow Cytometry

### Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

Sample preparation

For flow cytometry analysis of spleens, tissue was harvested, and dissociated into single cell suspension using syringe plungers and 40 micron cell filters. Red blood cells were lysed using ACK lysis buffer prior to staining. For flow cytometry of tumor infiltrating cells, tumors were harvested, manually disassociated into small pieces and digested with collagenase 1A. When appropriate, red blood cells were lysed prior to staining. For intracellular stains, cells were fixed and permeabilized using the BD Cytofix/Cytoperm Plus kit.

Instrument

Flow cytometry data was collected on either a BDFacs Canto II or BD LSRFortessa X-20.

Software

Flow cytometry data was collected using BD FACS Diva software version 8.0. Data was analyzed using FlowJo software version 10.2

Cell population abundance

No cell sorting was used.

Gating strategy

For MHC class I and class II tetramer staining, positive and negative population gates were defined based on staining with tumor-irrelevant tetramers. For, T-bet, Foxp3, PD-1, CD44, ICOS, CD150/SLAM, KLRG1, and iNOS staining, positive and negative populations were defined based on staining with an IgG antibody.

- ☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

# Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space

<https://doi.org/10.1038/s41586-019-1639-8>

Received: 24 October 2018

Accepted: 5 September 2019

Published online: 23 October 2019

Conor J. McClune<sup>1,2</sup>, Aurora Alvarez-Buylla<sup>1</sup>, Christopher A. Voigt<sup>2</sup> & Michael T. Laub<sup>1,3\*</sup>

Gene duplication is a common and powerful mechanism by which cells create new signalling pathways<sup>1,2</sup>, but recently duplicated proteins typically must become insulated from each other and from other paralogues to prevent unwanted crosstalk<sup>3</sup>. A similar challenge arises when new sensors or synthetic signalling pathways are engineered within cells or transferred between genomes. How easily new pathways can be introduced into cells depends on the density and distribution of paralogous pathways in the sequence space that is defined by their specificity-determining residues<sup>4,5</sup>. Here we directly investigate how crowded this sequence space is, by generating novel two-component signalling proteins in *Escherichia coli* using cell sorting coupled to deep sequencing to analyse large libraries designed on the basis of coevolutionary patterns. We produce 58 insulated pathways comprising functional kinase–substrate pairs that have different specificities than their parent proteins, and demonstrate that several of these new pairs are orthogonal to all 27 paralogous pathways in *E. coli*. Additionally, from the kinase–substrate pairs generated, we identify sets consisting of six pairs that are mutually orthogonal to each other, which considerably increases the two-component signalling capacity of *E. coli*. These results indicate that sequence space is not densely occupied. The relative sparsity of paralogues in sequence space suggests that new insulated pathways can arise easily during evolution, or be designed de novo. We demonstrate the latter by engineering a signalling pathway in *E. coli* that responds to a plant cytokinin, without crosstalk to extant pathways. Our work also demonstrates how coevolution-guided mutagenesis and the mapping of sequence space can be used to design large sets of orthogonal protein–protein interactions.

Many promising therapies, such as chimeric antigen receptor T cells<sup>6</sup> and engineered probiotic microorganisms<sup>7</sup>, require the ability to repurpose and transfer signalling pathways into new genomic contexts<sup>8–10</sup>. Similarly, new pathways arise during evolution through the duplication and diversification of existing signalling mechanisms. For engineered and evolved signalling proteins to execute independent functions within cells, they must minimize detrimental crosstalk, which is a substantial challenge for proteins with paralogues.

For signalling proteins such as protein kinases and their substrates, specificity is enforced primarily at the amino acid level<sup>11,12</sup>. These specificity-determining residues define a finite sequence space. How cells globally organize paralogous protein families in sequence space, and whether the specificity-determining residues of individual members of these families have been optimally distributed to minimize cross-talk during evolution, remain open questions. Work with SH3 and PDZ domains in eukaryotes has suggested that paralogues are densely packed in sequence space, a result of intense negative selection against crosstalk

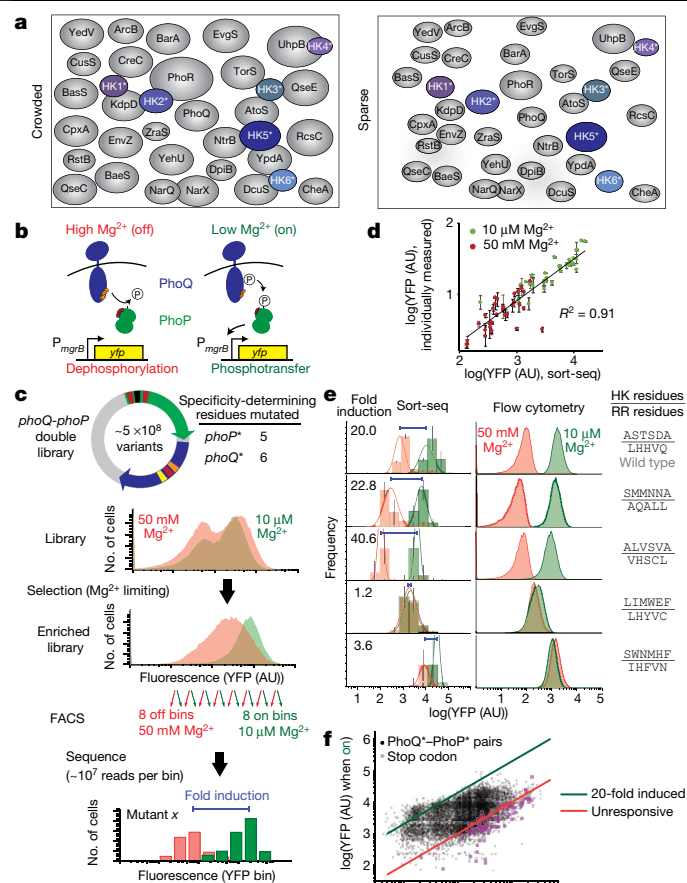
leading to a global optimization of specificity<sup>4,5</sup>. However, sequence space is vast and nature may not have fully occupied or explored it.

To assess how crowded paralogues are in sequence space, we sought to engineer protein complexes that are functional but insulated from extant paralogues. If sequence space is densely occupied by existing paralogues, it should be difficult to introduce new insulated pathways (Fig. 1a, Extended Data Fig. 1a). However, if sequence space is sparsely occupied, new pathways should be easy to introduce, and have a low probability of crosstalk. The design of orthogonal interacting proteins remains a major challenge, and previous efforts have generated only 3–4 orthogonal pairs<sup>13–15</sup>.

We focused on bacterial two-component signalling proteins, which involve a sensor histidine kinase that—upon activation—autophosphorylates and then transfers its phosphoryl group to a cognate response regulator to effect changes in cellular behaviour<sup>16</sup> (Fig. 1b, Extended Data Fig. 1b, c). Most histidine kinases are bifunctional, and act as phosphatases in the absence of a signal to dephosphorylate their cognate

<sup>1</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>2</sup>Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.

<sup>3</sup>Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA, USA. \*e-mail: laub@mit.edu



**Fig. 1 | Probing the density of paralogues in sequence space by building new orthogonal pairs of signalling proteins.** **a**, Two models for the distribution of paralogues in sequence space. Each oval is a niche that represents the set of substrates with which a kinase can interact, given its specificity-determining residues. These niches could be densely packed in sequence space (top) or more sparsely distributed (bottom), which would make the introduction of new insulated histidine kinases (denoted as HK1\* to HK6\*) relatively difficult or easy, respectively. **b**, Depending on  $Mg^{2+}$  levels, *E. coli* PhoQ (blue) can phosphorylate or dephosphorylate PhoP (green) to stimulate or repress, respectively, gene expression—including that of a YFP reporter (bottom). **c**, A library of about  $5 \times 10^8$  PhoQ–PhoP variants was first examined by flow cytometry before growth overnight in low levels of  $Mg^{2+}$ , followed by outgrowth with high or low levels of  $Mg^{2+}$ , and FACS, with deep sequencing of 8 consecutive bins. For each pair of variants, the number of reads in each bin was plotted to infer its fold-induction value. **d**, Pearson correlation between YFP values inferred by sort-seq, and measured individually. Points indicate mean  $\pm$  s.d.,  $n = 2$  biological replicates. AU, arbitrary units. **e**, For each pair of variants indicated (including the wild-type pair (top)), the signalling profile inferred by sort-seq (left) is compared to that measured in isolation. Experiment was repeated three times with similar results. RR, response regulator. **f**, Plot of mean YFP level inferred by sort-seq, for each pair of variants in the ‘off’ and ‘on’ states.

regulator. Bacteria usually encode dozens of two-component pathways (Fig. 1a, Extended Data Fig. 1d) that are mutually insulated; the vast majority of kinase–regulator pairs form exclusive, one-to-one relationships<sup>2</sup>. Both kinase and phosphatase activities, which involve the same protein–protein interface, contribute to pathway specificity<sup>11,17,18</sup>.

The specificity of interaction between kinases and regulators is driven by a limited set of interfacial residues in each protein that strongly coevolve with one another<sup>19–21</sup> (Extended Data Fig. 1c). To identify combinations of these residues that are functional and insulated from existing pathways in *E. coli*, we constructed a dual library of mutants in which the 11 key coevolving interface residues of a canonical two-component

system (PhoQ–PhoP) were randomized<sup>20,22</sup> (Fig. 1c, Extended Data Fig. 1c–e).

To identify functional combinations of residues, we first grew the library of PhoQ–PhoP variants overnight in medium with low  $Mg^{2+}$ , which activates PhoQ. Because cells must phosphorylate PhoP to grow when  $Mg^{2+}$  is limiting, this step enriches for functional PhoQ–PhoP variants (Fig. 1c). Variants that survived these limiting  $Mg^{2+}$  conditions were then subjected to fluorescence-activated cell sorting (FACS) and deep sequencing to quantify their signal responsiveness (sort-seq) (Extended Data Figs. 2, 3). To gauge PhoP activity in vivo, we used a fluorescent reporter,  $P_{mgrB}$ -yfp (Supplementary Tables 1, 2). In conditions of low extracellular  $Mg^{2+}$ , functional PhoQ promotes PhoP phosphorylation and production of YFP, whereas with high levels of  $Mg^{2+}$  PhoQ drives PhoP dephosphorylation to limit the accumulation of YFP (Fig. 1b).

To identify signal-responsive variants, we sorted cells from each condition into eight bins and deep-sequenced the randomized regions of variants in each bin (Fig. 1c). We calculated the frequency of each variant pair in each bin to yield distributions of individual variant pairs under conditions of low and high  $Mg^{2+}$ . We then assessed the mean levels of YFP in each condition, as well as the fold induction (or signal responsiveness) of each variant pair.

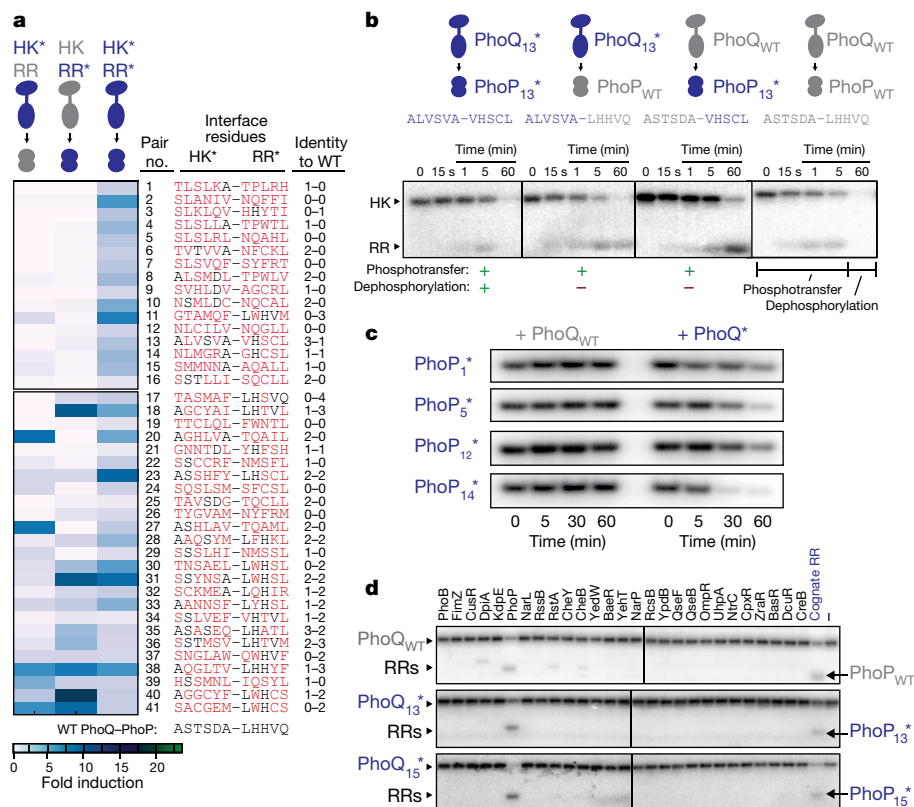
To validate our selection scheme, we isolated at random 48 individual clones from our starvation-enriched library and measured the distribution of YFP levels under conditions of low and high  $Mg^{2+}$  (Fig. 1d). Of these 48 clones, 32 had sufficient sequencing coverage for quantification by sort-seq. Each individual flow-cytometry profile showed a strong similarity to the distribution inferred from sort-seq, including the profiles of variants with high fold-induction values and those that were constitutively ‘on’ (Fig. 1e). The constitutively on behaviour probably arises when a PhoQ variant lacks phosphatase activity, as PhoP can then accumulate phosphoryl groups from other sources (for example, acetyl phosphate); this behaviour is seen with a  $\Delta$ phoQ strain<sup>21</sup> and in variants that contain a stop codon in *phoQ* (Fig. 1f).

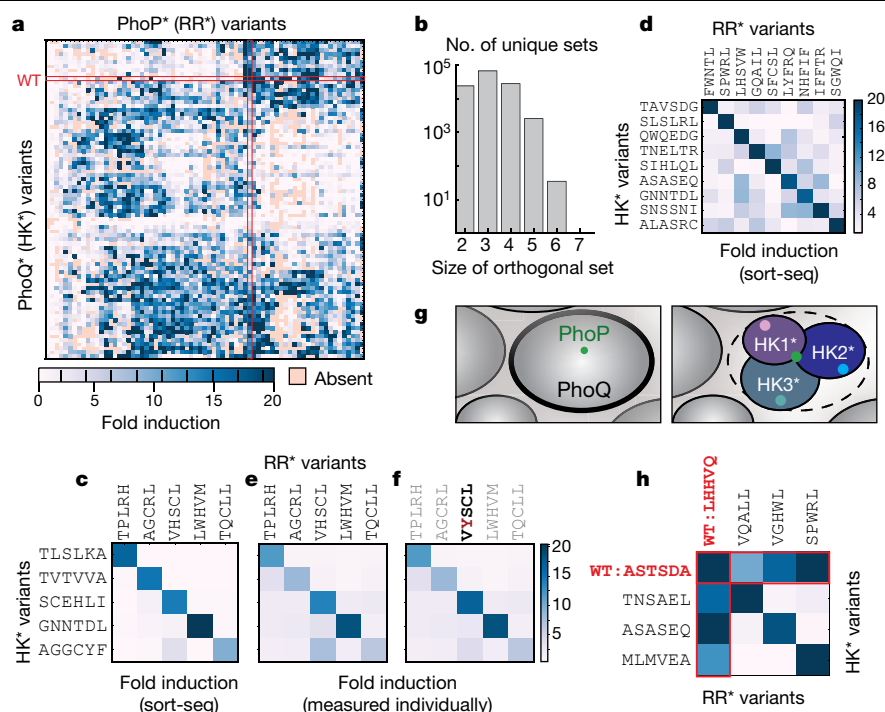
To select signal-responsive variants (similar to wild-type PhoQ–PhoP), we identified combinations of residues that produced fold-induction values of more than 20; there were 502 of these sequences, which we hereafter refer to as functional PhoQ\*–PhoP\* variants (these are numbered sequentially as PhoQ<sub>1</sub>\*–PhoP<sub>1</sub>\*, PhoQ<sub>2</sub>\*–PhoP<sub>2</sub>\* and so on) (Fig. 1f, Extended Data Fig. 2i). Most combinations of residues that were identified as functional shared few identities with the residues of wild-type PhoQ–PhoP. We isolated and characterized 41 diverse PhoQ\*–PhoP\* variants that showed more than 20-fold induction, and shared fewer than 5 residue identities with the wild-type PhoQ–PhoP at the 11 randomized positions (Fig. 2a). To test whether these PhoQ\*–PhoP\* pairs were insulated from the wild-type proteins, we built strains in which each of the PhoQ\* and PhoP\* variants was tested for interaction with a wild-type partner (Supplementary Table 3). For 16 of the 41 pairs we tested, there was substantially higher  $Mg^{2+}$ -dependent signalling with the mutant-protein pair than for either mutant paired with a wild-type partner (Fig. 2a). Thus, these 16 pairs do not crosstalk with the wild-type PhoQ–PhoP proteins, which indicates that our selection scheme produced functional signalling pathways that are insulated from the parental proteins.

To examine the mechanistic basis of insulation, we purified the PhoP\* variants and the cytoplasmic domains of the PhoQ\* variants for 7 of the 16 validated insulated protein pairs. The autokinase, phosphotransfer, and phosphatase activities of a histidine kinase can be assessed using a single assay<sup>11,16</sup> (Fig. 2b). Each kinase is autophosphorylated with [ $\gamma$ -<sup>32</sup>P] ATP and then mixed with a given partner, which results in phosphotransfer and a phosphorylated response regulator; as the unphosphorylated kinase accumulates it stimulates the dephosphorylation of a response regulator, which leads to the depletion of the radiolabelled response regulator.

For three of the functional mutant pairs, we observed the same pattern of activities as were observed with wild-type proteins, which







**Fig. 3 | Identification of sets of orthogonal signalling proteins.** **a**, Fold-induction values for all possible pairings from a set of 79 PhoQ\*–PhoP\* variants. The matrix was clustered in both dimensions. **b**, Number of unique sets of various sizes of orthogonal PhoQ\*–PhoP\* pairs, with fold-induction values >15 for cognate pairs and <6 for all non-cognate pairs. **c**, A set of five PhoQ\* and PhoP\* variants that are functional and mutually orthogonal. **d**, A set of 9 PhoQ\* and PhoP\* variants that are functional and mutually orthogonal, with fold-

induction values >17 for cognate pairs and <10 for non-cognate pairs. **e**, Fold-induction values for the mutant combinations in **c** measured individually. **f**, As in **e** but with a point mutant that reduces crosstalk (PhoP\*(VHSC) to PhoP\*(VYSCL)). **g**, Model for subfunctionalization of PhoQ in sequence space. **h**, Subfunctionalization of PhoQ specificity. A set of three PhoQ\*–PhoP\* variants that are mutually insulated but that retain interactions with the parent proteins.

identified are globally insulated from other pathways. Thus, there are unoccupied regions of sequence space, in which new systems with novel interaction specificities can be introduced without producing crosstalk to existing systems.

We also wanted to test the insulation of our selected functional PhoQ\*–PhoP\* variants with respect to each other. To this end, we selected 79 variant pairs that had high fold-induction values and a broad sequence diversity. We then combinatorially combined these PhoQ\* and PhoP\* variants, which produced a library with a theoretical diversity of 5,609. This library was transformed into cells containing the  $P_{merB}$ -yfp reporter, and was subjected to sort-seq (Fig. 1c, Extended Data Fig. 8a, b), which enabled us to infer the fold induction of each combination of PhoQ\* and PhoP\* variants. Within the resulting interaction matrix (Fig. 3a, Supplementary Table 4), 58 variant pairs were orthogonal to the wild-type proteins (Extended Data Fig. 8c).

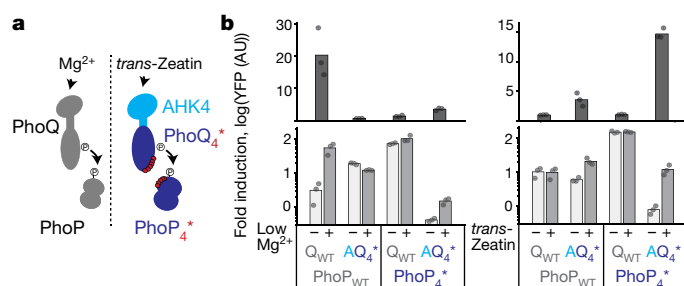
To isolate orthogonal sets of PhoQ\*–PhoP\* variants, we searched the 79 × 79 interaction matrix for sub-matrices in which strong interactions were seen only along the diagonal. We isolated more than 2,500 unique sets of 5 orthogonal signalling pairs, and dozens of sets of 6 orthogonal pairs (Fig. 3b, c, Extended Data Fig. 8d, e). With a slightly relaxed threshold for noncognate interactions, we found sets with up to 9 orthogonal protein pairs (Fig. 3d). To verify the orthogonality of these sets of PhoQ\*–PhoP\* variants, we cloned and analysed the 25 individual pairs that comprised a specific 5 × 5 matrix. Flow cytometry analysis showed strong agreement with the fold-induction values that were inferred using sort-seq (Fig. 3c, e).

Notably, the noncognate pairs in Fig. 3 were measured in the absence of the cognate partner of each variant. Any weak crosstalk that is seen should be eliminated by the phosphatase activity of the cognate kinase<sup>7,18</sup>; this prediction was confirmed for one instance of a PhoP\*

variant that exhibited modest crosstalk from wild-type PhoQ unless its cognate PhoQ\* was also expressed (Extended Data Fig. 8f, g). Thus, the off-diagonal values that are seen with the orthogonal sets in Fig. 3b–e represent upper limits on crosstalk. This small degree of crosstalk is also easily reduced further. For example, with the set in Fig. 3e, we screened point mutations of PhoP\*(VHSC) for reduced crosstalk, and found that PhoP\*(VYSCL) had reduced interaction with noncognate kinases while maintaining the interaction with its cognate kinase PhoQ\*(SCEHLI) (Fig. 3f).

The 79 × 79 matrix of interactions (Fig. 3a) also offered insight into how new pathways can arise through subfunctionalization (the partitioning of a niche in sequence space rather than movement into a new region) (Fig. 3g). For example, we found three pairs of PhoQ\*–PhoP\* variants that were insulated from each other, but each of which exhibited substantial crosstalk to the parental wild-type proteins (Fig. 3h, Extended Data Fig. 8h–k). Thus, these pairs have effectively partitioned the original niche of wild-type PhoQ (and PhoP) in sequence space to yield three insulated pathways. Subfunctionalization of duplicated proteins derived from a promiscuous ancestor may be a common mechanism through which insulated paralogs arise during evolution.

Collectively, our results indicate that the sequence space of two-component signalling pathways in *E. coli* is relatively sparsely occupied, such that new orthogonal signalling pathways can readily be introduced. The 502 functional variant pairs that we have isolated here have few specificity residues in common with wild-type PhoQ–PhoP, each other, or extant two-component signalling proteins (Extended Data Fig. 9a). A force-directed graph based on similarities in specificity residues highlights the diversity of naturally occurring interfaces and of the variants that we isolated (Extended Data Fig. 9b). To estimate how easily a new insulated pathway can be introduced, we noted that 502 functional



**Fig. 4 | Construction of an insulated sensor. a**, An insulated cytokinin (*trans*-zeatin) sensor (AQ<sub>4</sub>\*), constructed by fusing the sensory domain of *A. thaliana* AHK4 to PhoQ<sub>4</sub>\*. **b**, The chimeric sensor AQ<sub>4</sub>\* is specifically responsive to 1 μM *trans*-zeatin, phosphorylating its cognate mutant PhoP<sub>4</sub>\* to activate a YFP reporter. Wild-type PhoQ (Q<sub>WT</sub>) responds only to Mg<sup>2+</sup> and not to the cytokinin. Bars indicate mean from *n* = 3 biological replicates.

pairs came from 10,595 pairs with quantifiable fold-induction values. Of these 502, about 40% are probably insulated from wild-type PhoQ–PhoP (Fig. 2a), which implies that around 200 (1.6%) of the 10,595 pairs are both functional and insulated. This frequency is an upper bound, as the initial 10,595 pairs arose from selection under conditions of low Mg<sup>2+</sup> (Fig. 1c), which enriches about 100-fold for functionality (Extended Data Fig. 2b). Nevertheless, given the size of sequence space, these estimates underscore the relative ease of creating kinase–substrate pairs that are functional and orthogonal to their parent proteins.

Orthogonal signalling pathways will be useful in generating synthetic sensors and novel regulatory systems. As an example, we sought to generate a new pathway in *E. coli* that responds to the cytokinin *trans*-zeatin, a plant hormone. The histidine kinase AHK4 from *Arabidopsis thaliana* senses *trans*-zeatin, but crosstalks extensively with a native two-component pathway in *E. coli*<sup>23</sup>. To overcome this limitation, we fused the AHK4 sensory domain to the kinase domains of an orthogonal PhoQ\* and expressed this construct in *E. coli* with the cognate PhoP\* (Fig. 4a, b). This engineered sensor kinase enabled *E. coli* to respond specifically to *trans*-zeatin, and was insulated from all native two-component pathways, as measured by phosphotransfer profiling and RNA-seq (Extended Data Fig. 10). Thus, this chimeric sensor kinase and its cognate PhoP\* expand the sensory repertoire of *E. coli* without introducing undesirable crosstalk.

To date, synthetic circuits have mainly been built from nucleic-acid components because of their intrinsic modularity and programmability<sup>24</sup>. Protein-based circuits offer faster response times and richer functionality, but require more complicated programming of protein interactions. Our work enables the design of two-component signalling-based circuits in bacteria or eukaryotes, and the relatively sparse distribution of paralogues in sequence space means that multiple pathways can readily be introduced.

In summary, our work highlights the power of using coevolution-guided libraries to investigate protein–protein interactions and supports a model in which sequence space is not densely occupied. The relatively sparse distribution of extant proteins in sequence space presumably reflects their evolutionary history. A previous study indicated that duplicated signalling proteins are under pressure immediately after duplication to change and become insulated, but subsequent movement in sequence space then arises only from neutral changes<sup>3</sup>. Although

duplicated proteins are initially subject to selection against crosstalk with each other, each protein is probably not subject to system-wide negative selection or global optimization.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1639-8>.

- Alm, E., Huang, K. & Arkin, A. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLOS Comput. Biol.* **2**, e143 (2006).
- Capra, E. J. & Laub, M. T. Evolution of two-component signal transduction systems. *Annu. Rev. Microbiol.* **66**, 325–347 (2012).
- Capra, E. J., Perchuk, B. S., Skerker, J. M. & Laub, M. T. Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling families. *Cell* **150**, 222–232 (2012).
- Zarrinpar, A., Park, S. H. & Lim, W. A. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* **426**, 676–680 (2003).
- Stiffler, M. A. et al. PDZ domain binding selectivity is optimized across the mouse proteome. *Science* **317**, 364–369 (2007).
- Brentjens, R. J. et al. CD19-targeted T cells rapidly induce molecular remissions in adults with chemotherapy-refractory acute lymphoblastic leukemia. *Sci. Transl. Med.* **5**, 177ra38 (2013).
- Riglar, D. T. & Silver, P. A. Engineering bacteria for diagnostic and therapeutic applications. *Nat. Rev. Microbiol.* **16**, 214–225 (2018).
- Morsut, L. et al. Engineering customized cell sensing and response behaviors using synthetic Notch receptors. *Cell* **164**, 780–791 (2016).
- Bashor, C. J., Helman, N. C., Yan, S. & Lim, W. A. Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics. *Science* **319**, 1539–1543 (2008).
- Sockollosky, J. T. et al. Selective targeting of engineered T cells using orthogonal IL-2 cytokine-receptor complexes. *Science* **359**, 1037–1042 (2018).
- Skerker, J. M., Prasol, M. S., Perchuk, B. S., Biondi, E. G. & Laub, M. T. Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. *PLoS Biol.* **3**, e334 (2005).
- Creixell, P. et al. Unmasking determinants of specificity in the human kinome. *Cell* **163**, 187–201 (2015).
- Thompson, K. E., Bashor, C. J., Lim, W. A. & Keating, A. E. SYNZIP protein interaction toolbox: in vitro and in vivo specifications of heterospecific coiled-coil interaction domains. *ACS Synth. Biol.* **1**, 118–129 (2012).
- Boyken, S. E. et al. De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **352**, 680–687 (2016).
- Reinke, A. W., Grant, R. A. & Keating, A. E. A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering. *J. Am. Chem. Soc.* **132**, 6025–6031 (2010).
- Stock, A. M., Robinson, V. L. & Goudreau, P. N. Two-component signal transduction. *Annu. Rev. Biochem.* **69**, 183–215 (2000).
- Groban, E. S., Clarke, E. J., Salis, H. M., Miller, S. M. & Voigt, C. A. Kinetic buffering of cross talk between bacterial two-component sensors. *J. Mol. Biol.* **390**, 380–393 (2009).
- Siryaporn, A. & Goulian, M. Cross-talk suppression between the CpxA–CpxR and EnvZ–OmpR two-component systems in *E. coli*. *Mol. Microbiol.* **70**, 494–506 (2008).
- Capra, E. J. et al. Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways. *PLoS Genet.* **6**, e1001220 (2010).
- Skerker, J. M. et al. Rewiring the specificity of two-component signal transduction systems. *Cell* **133**, 1043–1054 (2008).
- Podgornaia, A. I. & Laub, M. T. Pervasive degeneracy and epistasis in a protein–protein interface. *Science* **347**, 673–677 (2015).
- Casino, P., Rubio, V. & Marina, A. Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. *Cell* **139**, 325–336 (2009).
- Yamada, H. et al. The *Arabidopsis* AHK4 histidine kinase is a cytokinin-binding receptor that transduces cytokinin signals across the membrane. *Plant Cell Physiol.* **42**, 1017–1023 (2001).
- Nielsen, A. A. et al. Genetic circuit design automation. *Science* **352**, aac7341 (2016).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### Bacterial strains and media

*E. coli* strains were grown in M9 medium (1× M9 salts, 100 μM CaCl<sub>2</sub>, 0.2% glucose, 0.1% casamino acids and MgSO<sub>4</sub> at indicated concentrations). When indicated, antibiotics were used at the following concentrations: carbenicillin 50 μg/ml, kanamycin 50 μg/ml, spectinomycin 50 μg/ml and chloramphenicol 32 μg/ml.

The base strain for all studies was *E. coli* strain TIM175 (MG1655  $\Delta$ *phoPQ* $\Delta$ *lacZYA* att $\lambda$ ::[*P<sub>mgrB</sub>*-*yfp*] attHK::[*P<sub>tetA</sub>*-*cfp+*])<sup>25</sup> with a ColEI/amp<sup>R</sup> plasmid (pCM150) containing *P<sub>mgrB</sub>*-*yfp*. All libraries were cloned onto a low-copy pSC101/spec<sup>R</sup> plasmid (pCM099, a derivative of pLPQ2), in which *phoP-phoQ* was driven by a constitutive *lacUV5* promoter. We also introduced a bicistron RBS (BCD18) upstream of *phoP-phoQ*<sup>26</sup>, which leads to expression of a single transcript that encodes a small (17 amino acid) open reading frame, followed by an independent ribosome binding site and then *phoP-phoQ*. This configuration ensures that mutations near the 5' end of the *phoP* coding region do not substantially affect expression by changing interactions between the 5' end of *phoP* and the upstream leader sequence. Expression from the *lacUV5* promoter in a plasmid probably produces more PhoQ–PhoP than is natively produced, which increases the chance of crosstalk; thus, the variant pairs identified as orthogonal would perform even better with respect to crosstalk at lower expression levels. Additional characterizations of PhoQ\* and PhoP\* variants isolated from the library were done with a three-plasmid setup: reporter plasmid pCM150, pCM143 (*lacUV5-BCD18-phoP*, pSC101/spec<sup>R</sup>) and pCM149 (*lacUV5-RBS\_B0034-phoQ*, p15A/kan<sup>R</sup>). Point mutations were introduced using blunt-end ligation<sup>27</sup> and Gibson assembly<sup>28</sup>.

### Flow cytometry characterization

To induce PhoQ–PhoP, cells were grown to mid-exponential phase (optical density at 600 nm (OD<sub>600</sub>) of about 0.5) in M9 before being washed once with M9 containing 0 mM MgSO<sub>4</sub> and diluted 1:100 into M9 containing 10 μM MgSO<sub>4</sub> (for induction) or 50 mM MgSO<sub>4</sub> (for repression). Cells were grown for 6 h, diluted 1:50 into PBS with 0.5 g/l kanamycin, and fluorescence was measured on a Miltenyi MACSQuant VYB. An identical procedure was used to induce AHK4–PhoQ fusions, except that cells were grown in M9 containing 2 mM MgSO<sub>4</sub> and 1 nM anhydrotetracycline (aTc) (Sigma) at all times, with the on condition containing 1 μM *trans*-zeatin (Sigma) and the off condition containing no *trans*-zeatin. In each cytometry experiment, three replicates of each sample were induced independently and 20,000 cells were measured per replicate. FlowJo was used to analyse the data, gating on single live cells and extracting the geometric mean of the YFP distribution (Extended Data Fig. 2j). Error bars indicate the s.d. of the geometric means measured in each replicate.

### Design and assembly of the degenerate PhoQ–PhoP library

The PhoQ–PhoP saturation mutation library was constructed by replacing the targeted residues with NNS codons<sup>29–31</sup>. The residues targeted were selected on the basis of amino acid coevolution analyses performed using GREMLIN<sup>32</sup>.

The plasmid library was assembled in two general steps: (1) individual PhoP and PhoQ libraries were built in separate vectors and (2) sections of these vectors were combined to produce a new vector containing both the PhoP and the PhoQ mutant (Extended Data Fig. 2a). For the first step, oligonucleotide libraries for the sections of *phoP* and *phoQ* to be mutated were ordered from DNA 2.0. NNS nucleotides replaced codons 12, 14, 15, 18 and 19 in PhoP and codons 284, 288, 289, 292, 302 and 303 in PhoQ. These oligonucleotides were cloned into vectors pCM0715 and pCM076 using the type IIS restriction enzyme BsmBI. A toxic *ccdB* locus on these plasmids (used as a counter selection) was replaced during the process, ensuring a high rate of insertion incorporation. Both insert and vector were digested with BsmBI at 55 °C for 2 h and then purified on a Zymo DNA clean column. Then, 1 pmol of both insert and vector was combined in a 25-μl reaction with 400 units of T4 ligase and incubated at 16 °C

for 16 h. Three ligations of each library were done, to ensure sufficient numbers of transformants. Ligations were dialysed on Millipore VSWP 0.025-μm membrane filters for 60 min and then the entire volume was electroporated into 20 μl of Invitrogen MegaX DH10B cells. From three ligations of each library, a total of  $2.3 \times 10^8$  and  $7.4 \times 10^7$  transformants were obtained for the PhoQ and PhoP libraries, respectively.

BsaI sites were used to join the two sublibraries into a single plasmid. The fusion points were designed such that faulty assemblies would not be viable: one junction was within the spec<sup>R</sup> cassette (both the PhoP and PhoQ sublibraries contained a kan<sup>R</sup> cassette, but each contained only one half of the spec<sup>R</sup> open reading frame) and the other was within PhoQ. In brief, 500 fmol of miniprep DNA product (Qiagen) from each of the two libraries was combined in 25 μl T4 ligase buffer and digested with 1 μl BsaI for 1 h at 37 °C. T4 ligase was then added and the reaction was cycled between 16 °C (3 min) and 37 °C (2 min) for 50 cycles to allow iterative ligation and digestion, running the reaction to completion. The final ligation product was dialysed on Millipore VSWP 0.025-μm membrane filters for 60 min and the entire volume was electroporated into 20 μl of Invitrogen MegaX DH10B cells. In total, 12 ligations and electroporations were done to produce a total of  $5.72 \times 10^8$  transformants. Transformations were pooled and grown overnight (14 h) in 100 ml 2× YT + carbenicillin and spectinomycin. Following assembly, the plasmid library was purified by miniprep (Qiagen), dialysed and electroporated into  $\Delta$ *phoP-phoQ* strain CJM2044 to yield  $3.8 \times 10^9$  transformants.

### Library selection and sort-seq

The PhoQ–PhoP library was subjected to an initial selection of Mg<sup>2+</sup> starvation to enrich for functional variants before performing FACS. To this end, 6 ml of an overnight culture of the library (in 2× YT) was washed in M9 and diluted to an OD<sub>600</sub> of about 0.1 in 100 ml M9 containing 2 mM MgSO<sub>4</sub>. Three replicates of this culture were made, and carried separately through the subsequent selection, FACS and deep-sequencing process. Cells were grown for approximately 2 h to OD<sub>600</sub> of about 0.4, at which point 1.6 ml of culture was washed 3 times in M9 containing no MgSO<sub>4</sub>, and used to inoculate 100 ml of M9 containing no MgSO<sub>4</sub>. After each dilution, the culture was sampled and a dilution series was plated on LB plates to ensure no bottlenecks occurred (colony-forming units >1 × 10<sup>9</sup>). The cultures in M9 containing no MgSO<sub>4</sub> were grown overnight (14 h), with the OD<sub>600</sub> increasing from 0.05 to only about 0.07. MgSO<sub>4</sub> was then added to bring the concentration to 2 mM, and cells were grown to an OD<sub>600</sub> of 0.5 in 6 h, at which point glycerol stocks were made.

For FACS, 1-ml glycerol stocks were thawed and inoculated in 25 ml of M9. For each library replicate, 1 frozen stock aliquot was added directly to M9 containing 50 mM MgSO<sub>4</sub> (off state) and 1 aliquot was washed 3 times in M9 containing 0 mM MgSO<sub>4</sub> before inoculation in M9 with 10 μM MgSO<sub>4</sub> (on state). To maintain cells in exponential phase, cultures were diluted (1:4 for on state, 1:10 for off state) after 3 h. After 6 h, cells were diluted again (1:5), and chloramphenicol was added to a concentration of 320 μg/ml and cells were placed on ice for sorting. CFP was expressed at a low constitutive level (attHK::[*P<sub>tetA</sub>*-*cfp*]) and used to normalize YFP expression. Cells were sorted into ratiometric bins on the diagonal of CFP and YFP expression, to control for extrinsic expression noise in the YFP signal. For each library replicate, both the on and off cultures were sorted into 8 separate bins, generating 48 total bins. Up to 2.5 million cells were sorted into bins per replicate (Extended Data Figs. 2c, 8a). Sorted cells were added to 2× YT medium containing 2 mM MgSO<sub>4</sub>, carbenicillin and spectinomycin, and then grown overnight.

### Illumina sample preparation

After FACS, plasmids were purified (Qiagen MiniPrep) from overnight cultures representing each bin from each library replicate. For the two mutagenized regions of the plasmid to be brought into close enough proximity (<790 bp) for paired-end Illumina sequencing (Extended Data Fig. 2b), plasmids were digested with XhoI and then self-ligated (T4 ligase, 4 h). To isolate only self-ligation products and not cross-ligation



products, ligation reactions were cleaned (Zymo PCR Clean Up) and gel-purified to select for the correct size on FlashGels (Lonza). Two PCR reactions were performed, both using KAPA HiFi Hotstart, to add Illumina sequencing adaptors and barcodes. First, ligation reaction products were amplified for 30 cycles (95 °C for 30 s, 65 °C for 15 s and 72 °C for 120 s) with Illumina inner amplification primers (Supplementary Table 3) in an emulsion PCR (Micellula Emulsion PCR) to avoid PCR chimeras. Second, purified PCR product from the first reaction was subjected to a second PCR with barcoding primers for 9 cycles (95 °C for 30 s, 65 °C for 15 s and 72 °C for 60 s). Final products were quantified (NanoDrop), normalized, combined and sequenced on an Illumina NextSeq. For each bin, 1–33 million reads were collected.

## Construction of combinatorial mutant library

Seventy-nine pairs of PhoQ\*–PhoP\* variants were selected that displayed broad sequence diversity and high fold induction in the initial PhoQ\*–PhoP\* library. These pairs included 79 unique PhoQ\* variants and 71 unique PhoP\* variants. These unique variants were cloned into plasmids pCM143 (PhoP) and pCM149 (PhoQ) using blunt-end ligation<sup>27</sup> and Gibson assembly<sup>28</sup>. Each pCM143 and pCM149 variant was amplified by PCR (KAPA HiFi, 30 cycles) to generate amplicons for Gibson assembly (primers CM937/CM1531 for pCM149 and primers CM938/CM1532 for pCM143). PCR products were cleaned (Zymo PCR CleanUp), quantified by NanoDrop and combined into an equimolar mix of pCM143 amplicons and an equimolar mix of pCM149 amplicons. The two mixes were combined in Gibson assembly master mix (300 fmol of large pCM143 fragment and 900 fmol of the smaller pCM149 insert), incubated at 50 °C for 2 h and heat-killed at 79 °C for 20 min. The assembly was dialysed on Millipore VSWP 0.025-μm membrane filters for 60 min and transformed into electrocompetent CJM2044 cells.

Unlike the treatment of the initial larger library, this library was not subjected to a low-magnesium selection step. Immediately after construction, this library underwent sort-seq, as described in ‘Library selection and sort-seq’.

## Illumina data processing

The frequency of each mutant in each bin was calculated by taking the fraction of reads in a given bin that corresponded to a given sequence, normalized by the fraction of cells in that given bin (Supplementary Information). All sort-seq plots display the mean frequencies in each bin across three replicates, and error bars indicate the s.d. Gaussian functions were fit to each distribution (in log<sub>10</sub>(YFP units)), from both the on and off sorts (SciPy optimize package). Variants with fewer than 25 total reads were discarded before fitting. Poor Gaussian fits have high variances on the estimated parameters. The s.d. error on the estimated log(YFP) mean ( $\sigma_{\text{fit}}$ ) was used as a metric to filter poorly fit sequences: sequences were removed if  $\sigma_{\text{fit, on}} + \sigma_{\text{fit, off}} > 2$ . In total, 10,595 unique variants passed these filters. Fold-induction values were calculated as the ratio of the fit means between the induced and uninduced states:  $\mu_{\text{on}}/\mu_{\text{off}}$ .

During the analysis of the second, combinatorial library (Fig. 3a), the fold induction was calculated for the most-frequent nucleotide sequence that represented each amino acid sequence. For visualization and orthogonal set design, fold inductions were bounded between 1 and 20. Individually tested mutants generally did not surpass the sensitivity of wild type PhoQ–PhoP, which displayed 20-fold induction during sort-seq; this suggests that signal above 20-fold may be due to noise. The axes of the matrix in Fig. 3a were clustered hierarchically using the WGPCM method (Scipy). During clustering, matrix entries that lacked data were assigned the mean value of all other entries.

## Analysis of the sensitivity of sort-seq quantification to read count

To assess the quality of sort-seq-based quantification for variants with lower read coverage, variants with high read coverage (2,000–10,000 total reads) were downsampled to simulate low read coverage and fed through the sort-seq analysis pipeline (Extended Data Fig. 3). For each of

these high-coverage variants, simulated data were produced by down-sampling 100 independent times by the factors indicated in Extended Data Fig. 3. Simulated read coverage was generated by sampling (with replacement) from the original reads of each variant up to the desired read coverage. Simulated reads were then subjected to the same Gaussian fitting protocol as described in ‘Illumina data processing’. As in the original analysis, poor fits ( $\sigma_{\text{fit, on}} + \sigma_{\text{fit, off}} > 2$ ) were discarded. All simulated variants were classified as functional or non-functional, on the basis of the fold-induction values of the original high-coverage variants from which they were sampled (Extended Data Fig. 3c, e). False-positive rates (Extended Data Fig. 3d) and false-negative rates (Extended Data Fig. 3f) were then calculated as a function of read coverage by computing the fraction of simulated variants that were misclassified.

## Orthogonal set design

Orthogonal sets of PhoQ\* and PhoP\* variants of up to seven pairs in size were identified by systematically scanning all PhoQ\*–PhoP\* permutations within the matrix in Fig. 3a and Supplementary Table 4. Larger orthogonal sets were identified using a greedy search algorithm (Supplementary Information). The 5 × 5 orthogonal set described in Fig. 4e was further optimized by testing single-point mutants of the single PhoP\* variant (VHSCL) that initially displayed crosstalk with a noncognate PhoQ\* variant. Each of the five PhoP\* specificity residues was replaced independently with an NNK codon to generate all possible single-point mutants (XHSCL, VXSC, VHXSL, VHSXL and VHSCX, in which X is any amino acid specified by the NNK codon). These mutants were cotransformed with the cognate PhoQ\* variant (SCEHLI) into CJM2044 and grown overnight in M9 medium containing 0 mM MgSO<sub>4</sub> to remove nonfunctional variants. After plating the surviving strains on LB agarose plates, 48 clones were tested for Mg<sup>2+</sup> induction (see ‘Flow cytometry characterization’). The pCM143-PhoP\* plasmids from the 24 clones with the strongest induction were purified and cotransformed with the noncognate PhoQ\* variant (AGGCYF) into CJM2044. Mg<sup>2+</sup> induction was measured by cytometry and the eight clones displaying the highest cognate/noncognate induction ratio were selected for testing with all five PhoQ\* variants. Two of these eight clones were PhoP\*(VYSCL), which displayed the highest specificity.

## Reconstruction and in vivo characterization of individual PhoQ\* and PhoP\* variants

Variants were cloned into plasmids pCM143 (PhoP) and pCM149 (PhoQ) using blunt-end ligation<sup>27</sup> and Gibson assembly<sup>28</sup>. Combinations of pCM143 and pCM149 plasmids were cotransformed into strain CM2044. Colonies were grown overnight in M9 containing 2 mM MgSO<sub>4</sub> and induced as described in ‘Flow cytometry characterization’ with M9 containing either 10 μM or 50 mM MgSO<sub>4</sub>. After 6 h, cultures were diluted 1:50 into cold PBS containing 0.5 g/l kanamycin, and fluorescence was measured on a Miltenyi MACSQuant VYB. The fold-induction values of individually tested variant pairs were generally smaller than those measured by sort-seq, probably owing to differences between the Miltenyi cytometer and BD Aria sorter; however, the two measurements were highly correlated (Pearson  $R^2 = 0.91$ ) (Fig. 1d).

## Purification of two-component signalling proteins and in vitro phosphotransfer assays

Expression and purification of PhoQ\* and PhoP\* variants, and phosphotransfer experiments, were carried out as previously described<sup>11,21</sup>. PhoP\* was purified fused to a His<sub>6</sub>-Trx tag, and the cytoplasmic region of PhoQ\* (residues 238–486) was fused to a His<sub>6</sub>-MBP (maltose-binding protein) tag. For phosphotransfer reactions, the kinase was autophosphorylated for 1 h at 30 °C with [ $\gamma$ -<sup>32</sup>P]ATP (Perkin Elmer) before being combined with PhoP\* at a 1:8 ratio (10-μl reactions contained 1 μM PhoQ\* and 8 μM PhoP\*). Reactions were stopped at appropriate times by adding 4× Laemmli buffer with 8% 2-mercaptoethanol. This process enabled monitoring of both phosphotransfer and phosphatase activities between PhoQ\* and PhoP\* variants (Fig. 2b).



For PhoQ\* variants for which in vitro autophosphorylation was not observed, phosphatase activity was assayed by mixing a given PhoQ\* variant with a PhoP\* variant that was phosphorylated using wild-type PhoQ. This was achieved by incubating 8  $\mu$ M PhoP\* for 1 h at 30 °C with [ $\gamma$ - $^{32}$ P]ATP and 1  $\mu$ M wild-type PhoQ, which promiscuously phosphorylates—but does not dephosphorylate—most PhoP variants. After generating phosphorylated PhoP\*, 1  $\mu$ M of the PhoQ\* variant was added and samples taken and reactions stopped.

Phosphatase activity of PhoQ\* variants with respect to other response regulators was measured with a similar assay. Twelve *E. coli* response regulators were selected for their ability to be stably phosphorylated in vitro by a cocktail of six *E. coli* histidine kinases (CreC, RstA, PhoR, PhoP, EnvZ and CpxA, each at 250 nM). After 2 h of pre-incubation with radiolabelled ATP and this kinase cocktail, each regulator was combined with 2  $\mu$ M PhoQ or PhoQ\* variant. Reactions were stopped at 0, 60 and 120 min by adding 4 $\times$  Laemmli buffer with 8% 2-mercaptoethanol.

Response regulators for phosphotransfer profiles were purified as done for PhoP\*. Each was fused to an N-terminal His<sub>6</sub>-Trx tag, expressed in BL21(DE3) cells and purified on a Ni<sup>2+</sup>-NTA column<sup>3</sup>. Conditions for phosphotransfer profiles were also identical to above conditions; 10  $\mu$ l reactions containing 1  $\mu$ M [ $\gamma$ - $^{32}$ P] autophosphorylated PhoQ\* and 8  $\mu$ M response regulator were generated and stopped after 5 or 60 min. Gel images were analysed using quantified with ImageJ.

### RNA-seq

Cultures were grown overnight in M9 medium containing 2 mM MgSO<sub>4</sub> and then diluted 1:25 into fresh M9 containing 2 mM MgSO<sub>4</sub> and grown for 2 h to reach OD<sub>600</sub> of about 0.5. For the off condition, 1 ml of cells was diluted into 2 ml M9 containing 74 mM MgSO<sub>4</sub> for a final concentration of 50 mM MgSO<sub>4</sub>. For the on condition, 2 ml of cells were pelleted, washed twice with M9 + 10  $\mu$ M MgSO<sub>4</sub>, resuspended in M9 + 10  $\mu$ M MgSO<sub>4</sub>, and then 1 ml of cells was diluted into 2 ml M9 + 10  $\mu$ M MgSO<sub>4</sub>. The induction of AQ<sub>4</sub>\* strains was identical, except that cells were induced for 30 min in M9 (2 mM MgSO<sub>4</sub>) containing either 0  $\mu$ M *trans*-zeatin (off condition) or 1  $\mu$ M *trans*-zeatin (on condition).

RNA was collected as previously described<sup>33</sup>. After 30 min of growth, cells from each condition were collected by adding 1.8 ml culture to 200  $\mu$ M cold stop solution (95% ethanol, 5% acid buffered phenol, 4 °C). The mixture was centrifuged for 30 s at 13,000 r.p.m. on a benchtop centrifuge, and the supernatant was removed and the pellet flash-frozen in liquid nitrogen and stored at -80 °C. To extract RNA, TRIzol (Invitrogen) was heated to 65 °C, added directly to the pellet and incubated at 65 °C for 10 min with shaking at 2,000 r.p.m. (Eppendorf Thermomixer). The mixture was frozen at -80 °C for at least 10 min. After thawing, cells were centrifuged at 15,000 r.p.m., 4 °C for 5 min and the supernatant was removed into 400  $\mu$ l ethanol. The mixture was applied to a Direct-Zol spin column (Zymo) and centrifuged for 30 s at 13,000 r.p.m. The columns were washed with Direct-Zol RNA prewash buffer twice (400  $\mu$ l) and RNA wash buffer (700  $\mu$ l) once before eluting in 90  $\mu$ l DEPC water. Then 10  $\mu$ l of 10 $\times$  Turbo DNase buffer and 2  $\mu$ l Turbo DNase (Invitrogen) were added to the eluant. The mixture was digested at 37 °C for 20 min, followed by the addition of 2  $\mu$ l DNase and another 20 min incubation. The total volume was brought to 200  $\mu$ l with DEPC water and combined with 200  $\mu$ l acid-phenol:chloroform (IAA, Invitrogen), vortexed and centrifuged for 10 min at 21,000g and 4 °C. The top (aqueous) layer was extracted and ethanol was precipitated in 20  $\mu$ l NaOAc (3M), 2  $\mu$ l GlycoBlue (Invitrogen) and 600  $\mu$ l cold ethanol. Precipitation mix was incubated at -80 °C for more than 4 h before centrifuging for 30 min at 21,000g and 4 °C. The pellet was washed twice with 500  $\mu$ l cold 70% ethanol, then air-dried and resuspended in 50  $\mu$ l DEPC water. RNA integrity was validated on a 6% TBE-urea acrylamide Novex gel (Invitrogen) and yield was quantified by NanoDrop spectrophotometer. Ribosomal RNA was removed with the RiboZero rRNA Removal Kit for Bacteria (Illumina). RNA was fragmented and cDNA libraries were prepared at the MIT BioMicro Center sequencing core using the KAPA RNA HyperPrep

Kit (Roche) and sequenced on an Illumina HiSeq. Reads were mapped to the *E. coli* genome and plasmids with bowtie2 using default parameters<sup>34</sup>.

To determine whether selected PhoQ\*-PhoP\* variants interfered with other two-component signalling pathways, we examined whether any other response regulator or histidine kinase genes were upregulated transcriptionally when PhoQ\* variants were activated by low levels of Mg<sup>2+</sup>. We calculated the fold change in expression of each two-component regulatory gene as a ratio of reads under conditions of low and high Mg<sup>2+</sup> (Extended Data Fig. 7b). *rstA* and *rstB* are part of the PhoP regulon, and are directly upregulated by wild-type PhoQ-PhoP and most variant pairs. To quantify how these fold changes compared to wild-type PhoQ-PhoP, we calculated the ratio of the fold change in each gene to the geometric mean of the fold change in the same gene for the two wild-type replicates (Extended Data Fig. 7c). To assess whether any of the two-component signalling genes were significantly upregulated, we calculated the z-score of each ratio of fold changes and conducted a one-tailed test to compute *P* values (Extended Data Fig. 7d). After using a Bonferroni correction for multiple hypothesis testing, no genes were found to be significantly upregulated (*P* < 0.05).

### Identification of two-component signalling proteins and generation of force-directed graphs

The RefSeq Prokaryotic Genomes database of 5,506 bacterial genomes (September 2017) was downloaded from NCBI. The database was scanned for histidine kinases and response regulators using jackhmmer<sup>35</sup> (*e* value cut-off = 0.01) with all two-component signalling proteins from *E. coli* used as queries. The combined lists of histidine kinase and response regulator hits were aligned with hmmalign<sup>35</sup> to the PFAM hidden Markov models for HisKA and Response\_Reg domain families, respectively. Columns in the multiple sequence alignment with greater than 80% gaps were eliminated, and sequences with greater than 50% gaps were discarded. Histidine kinases that lack the catalytic histidine and response regulators that lack the catalytic aspartate were removed. Proteins that contain both the histidine kinase DHP domain and a response regulator receiver domain were discarded to avoid ambiguity. Histidine kinases and response regulators were then labelled as exclusive pairs if they were (i) within 20 genes in the genome with no other histidine kinase or response regulator gene between, (ii) on the same strand and (iii) closer to no other potential histidine kinase or response regulator partner (with distance defined as the number of genes between partners). The sequences of paired histidine kinases and response regulators were concatenated and the multiple sequence alignment was then reduced to the 11 positions mutated in this study.

The force-directed graph was generated using the Gephi network visualization package<sup>36</sup>. To construct a network, the 85,782 co-operonic histidine kinase-response regulator pairs identified by HMMER in bacterial genomes were combined with the functional mutant sequences from the PhoQ-PhoP dual library that had fold-induction values >18 and the mutant variants within the characterized 5 $\times$ 5 orthogonal set (Fig. 3c, e). These sequences were treated as nodes and were connected by edges if the pairwise alignment score for the 11 specificity residues (using the BLOSUM62 scoring matrix) exceeded a threshold score of 20. If more than 40 edges were connected to a node, only the top scoring 40 edges were kept. If no edge scoring above 20 connected a node that node retained its top-scoring edge, despite that edge being below the BLOSUM62 threshold. A final model of about 86,000 nodes and 2.5 million edges was loaded into Gephi and visualized using the Force-Atlas-2 tool<sup>36</sup>.

### Construction of AHK4-PhoQ chimera

Chimeric histidine kinases sensors were made using a variation of the 'primer aided truncation for the creation of hybrid proteins' (PATCHY) strategy<sup>37</sup>. The N-terminal region of AHK4 (residues 1-475) was cloned downstream of the P<sub>lac</sub> promoter on a p15A/kan<sup>R</sup> vector. This plasmid was amplified by PCR with primers containing SapI sites to enable insertion of the PhoQ kinase domain. Five sets of primers allowed five possible junction

sites within AHK4 (residues A466, A468, A469, A472 and A478) with identical GCG overhangs. PhoQ (pCM149) was amplified with 32 distinct primers (also containing SapI sites) to generate 32 C-terminal truncations beginning upstream of the DHP domain (residues 213–224 and 257–276). PCR products were gel-purified (Zymo), then combined in a 50- $\mu$ l ligation reaction containing 400 U of T4 ligase (NEB), 20 U SapI (NEB), 100 fmol pooled AHK4 PCR products and 500 fmol pooled PhoQ PCR products. The reaction was cycled 50 times between 37 °C (2 min) and 16 °C (3 min) to drive assembly to completion, heat-killed at 50 °C (20 min) and 80 °C (20 min) and dialysed on Millipore VSWP 0.025- $\mu$ m membrane filters (60 min). This small library of 160 possible fusions was transformed into electrocompetent CJM2044 cells containing PhoP on a plasmid (pCM143).

To enrich for chimeras responsive to *trans*-zeatin we used Mg<sup>2+</sup> starvation as a selection. An overnight culture of the library in M9 medium was induced for 1 h (M9 containing 21 nM aTc, 1  $\mu$ M *trans*-zeatin), washed three times in M9 containing no MgSO<sub>4</sub> and diluted 1:10 into 100 ml M9 containing no MgSO<sub>4</sub>, 21 nM aTc, 1  $\mu$ M *trans*-zeatin. After 4 h, 500  $\mu$ l of culture was plated on LB. Ninety-six colonies were picked and screened for *trans*-zeatin-dependent YFP expression.

## GCN4–DHP fusions to test phosphatase buffering against crosstalk in vivo

To generate cytosolic variants of PhoQ locked in a phosphatase state, we followed a previously described strategy<sup>38</sup> and fused GCN4 (MKQLED-KVEELLSKNYHLENEVARL) N-terminal to the DHP domain of PhoQ. pCM149 (*lacUV5-phoQ*, *kanR*, *p15A*) was amplified with 24 distinct primers (containing SapI sites) to generate 24 C-terminal truncations beginning upstream of the DHP domain (residues 222–225 and 257–276). The N terminus of PhoQ was replaced by GCN4 in each of these plasmids, removing the transmembrane and sensory domains. Each GCN4 fusion plasmid was transformed with pCM143 (*phoP*) and pCM150 (*P<sub>mgrB</sub>-yfp*) into TIM175 and tested by standard Mg<sup>2+</sup> induction (see ‘Flow cytometry characterization’) for activity. As expected, some variants displayed constitutive high YFP (presumably a locked kinase conformation) or constitutive low YFP (presumably a locked phosphatase conformation) and stepwise amino acid insertions displayed a periodicity of these phenotypes (Extended Data Fig. 8f). One of these fusions (fusion-266) displayed even lower constitutive YFP values than PhoQ with mutations in the ATP cap (R434M, R439M and Q442M; pCM180) or ATP pocket (N385L, N389L, K392M and Y393F; pCM179)<sup>39</sup>.

To test the ability of a cognate phosphatase to suppress crosstalk from a noncognate kinase, we used a three-plasmid setup: pCM874 (reporter plasmid pCM150 with PlacUV5-PhoP<sub>15</sub>\* inserted), pCM149 (PlacUV5-PhoQ<sub>WT</sub>) and pCM873 or pCM898 (Ptet-GCN4-fusion266-PhoQ<sub>WT</sub> or -PhoQ<sub>15</sub>\*, respectively). Because PhoP\* has been moved from a low-copy to medium-copy plasmid, crosstalk between PhoQ<sub>WT</sub> and PhoP\* is probably exacerbated by overexpression, as noted by the high level of induction seen in Extended Data Fig. 8g before aTc is added. However, induction of the GCN4–fusion-266–PhoQ<sub>15</sub>\* phosphatase effectively eliminates this crosstalk (Extended Data Fig. 8g, right). Induction of the noncognate phosphatase GCN4–fusion-266–PhoQ<sub>WT</sub>\* does not relieve this crosstalk.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

Datasets generated during this study have been deposited in GEO. Raw reads and processed sort-seq analysis of each mutant can be found under accession numbers GSE120780 (degenerate PhoQ–PhoP library) and GSE120786 (combinatorial library of 79 PhoQ\*–PhoP\* variants). Raw reads and reads per kilobase of transcript per million mapped reads (RPKM) for all *E. coli* genes from RNA-seq are deposited with accession number GSE128611.

## Code availability

Python scripts for analysis are available at <https://github.com/mcclune/nature2019>.

- Miyashiro, T. & Goulian, M. Stimulus-dependent differential regulation in the *Escherichia coli* PhoQ PhoP system. *Proc. Natl Acad. Sci. USA* **104**, 16305–16310 (2007).
- Mutalik, V. K. et al. Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–360 (2013).
- Ashenberg, O., Keating, A. E. & Laub, M. T. Helix bundle loops determine whether histidine kinases autophosphorylate in *cis* or in *trans*. *J. Mol. Biol.* **425**, 1198–1209 (2013).
- Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
- Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
- Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409–413 (2017).
- Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *eLife* **7**, e32472 (2018).
- Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S. I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).
- Culviner, P. H. & Laub, M. T. Global analysis of the *E. coli* toxin MazF reveals widespread cleavage of mRNA and the inhibition of rRNA maturation and ribosome biogenesis. *Mol. Cell* **70**, 868–880 (2018).
- Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
- Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
- Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE* **9**, e98679 (2014).
- Ohlendorf, R., Schumacher, C. H., Richter, F. & Möglich, A. Library-aided probing of linker determinants in hybrid photoreceptors. *ACS Synth. Biol.* **5**, 1117–1126 (2016).
- Wang, B., Zhao, A., Novick, R. P. & Muir, T. W. Activation and inhibition of the receptor histidine kinase AgrC occurs through opposite helical transduction motions. *Mol. Cell* **53**, 929–940 (2014).
- Marina, A., Mott, C., Auyzenberg, A., Hendrickson, W. A. & Waldburger, C. D. Structural and mutational analysis of the PhoQ histidine kinase catalytic domain. Insight into the reaction mechanism. *J. Biol. Chem.* **276**, 41182–41190 (2001).

**Acknowledgements** We thank I. Nocedal, P. Culviner, D. Ding and A. Podgornaia for helpful discussions. M.T.L. is an Investigator of the Howard Hughes Medical Institute. This work was also supported by a grant from the Office of Naval Research (N000141310074) to M.T.L. and C.A.V. and by the NIH pre-doctoral training grant T32GM007287.

**Author contributions** C.J.M., C.A.V. and M.T.L. conceptualized and designed the study. C.J.M. performed all experiments with assistance from A.A.-B. C.J.M., M.T.L. and C.A.V. wrote the manuscript. C.A.V. and M.T.L. supervised the study and provided funding support.

**Competing interests** The authors declare no competing interests.

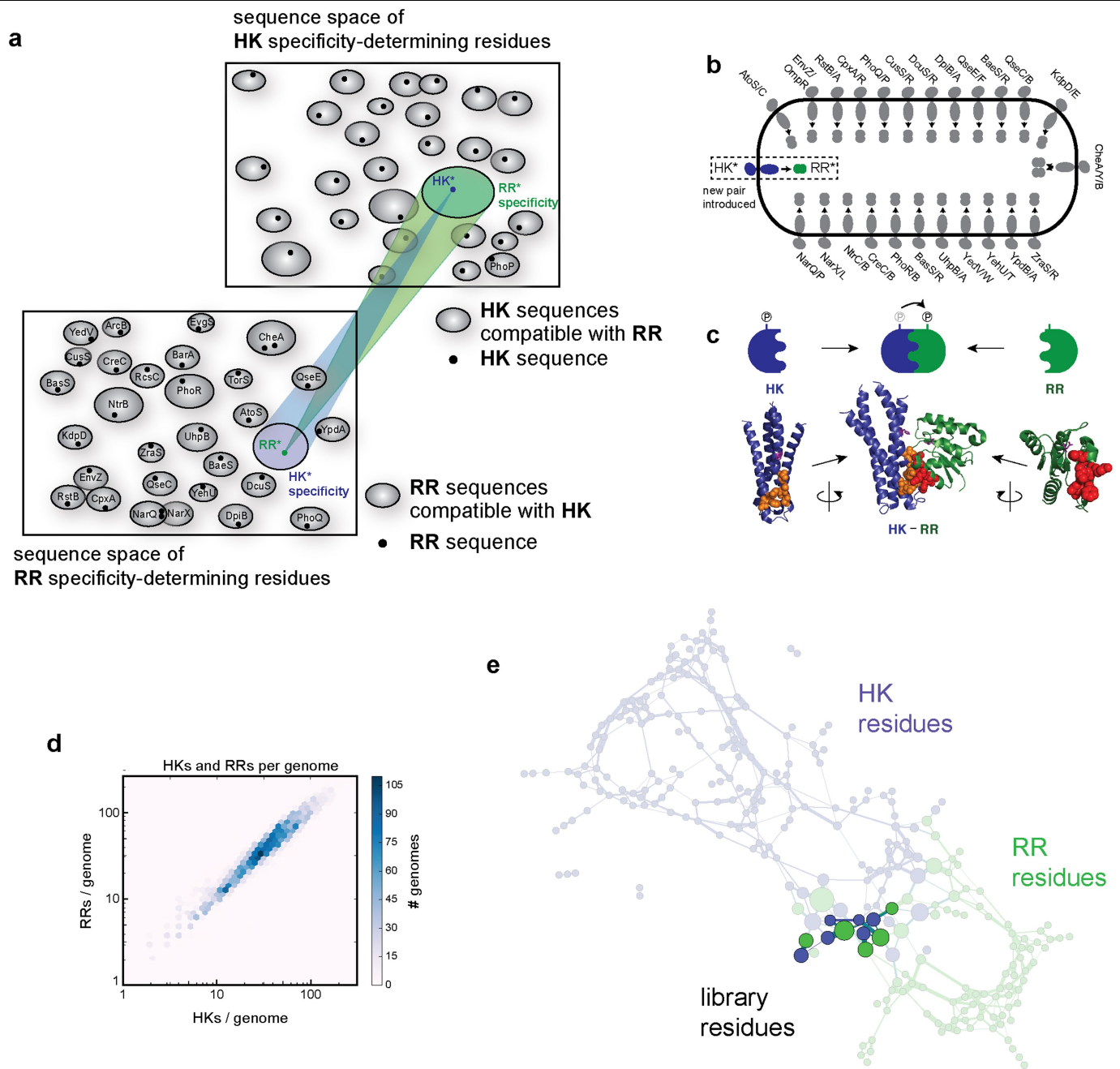
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1639-8>.

**Correspondence and requests for materials** should be addressed to M.T.L.

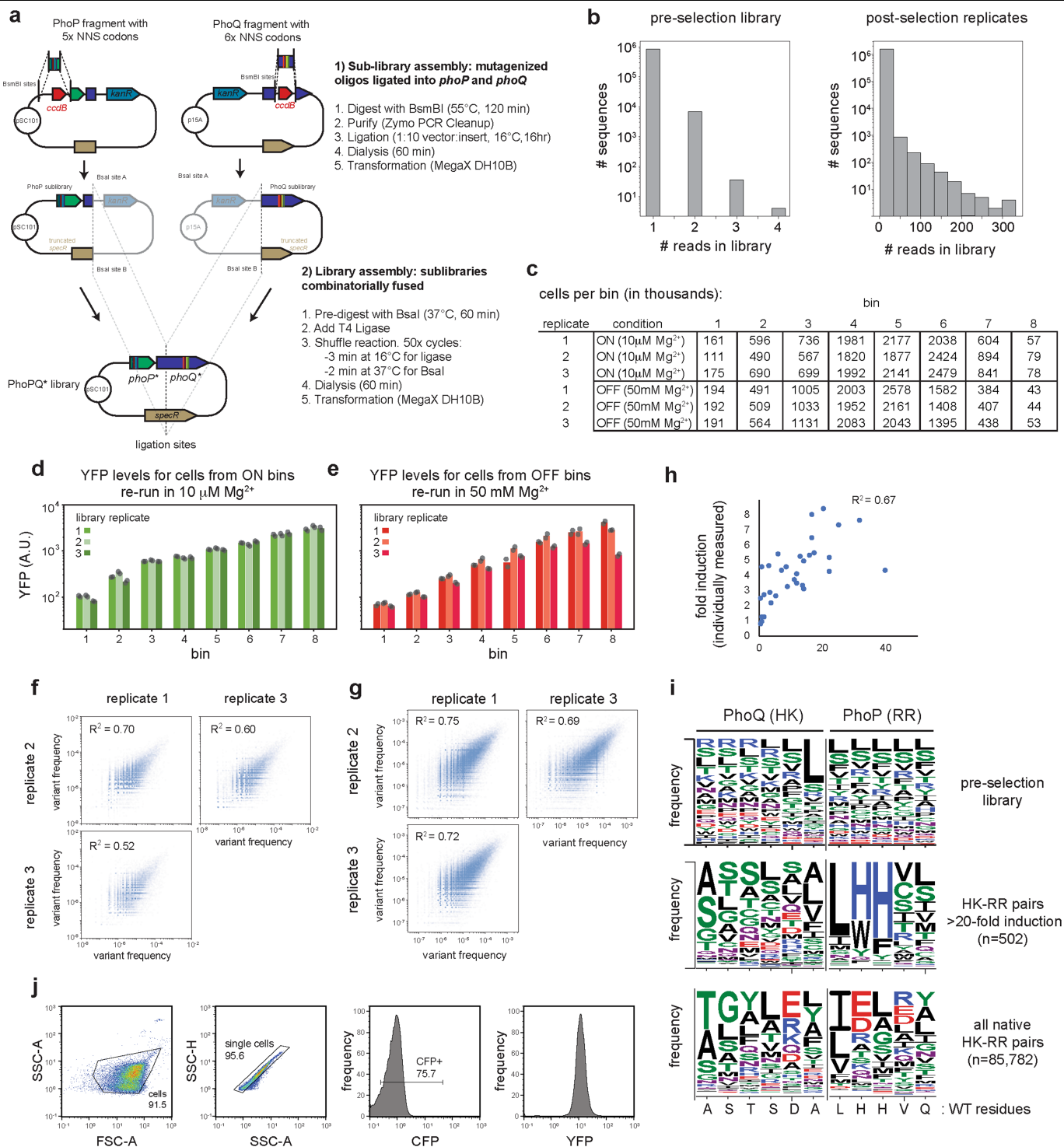
**Peer review information** Nature thanks Tamar Friedlander and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Bioinformatic and coevolutionary analysis of two-component signalling systems.** **a**, Schematic illustrating the challenge of identifying HK\*–RR\* pairs that are orthogonal to all endogenous histidine kinases and response regulators. For both histidine kinases and response regulators, the specificity-determining residues define a finite sequence space. The specificity-determining residues of each histidine kinase determine the set of response regulators with which it can interact. These sets, or niches in sequence space, are depicted as ovals, and each cognate response regulator is represented by a black dot (bottom left). A similar representation is shown for each response regulator and the set of histidine kinases with which it can interact (top right). The two sequence spaces are connected, as depicted with coloured cones for a single histidine kinase–response regulator pair. The establishment of a new signalling pathway that is orthogonal to existing systems requires that the two new proteins are compatible with each other, but occupy regions of histidine-kinase and response-regulator specificity space that are

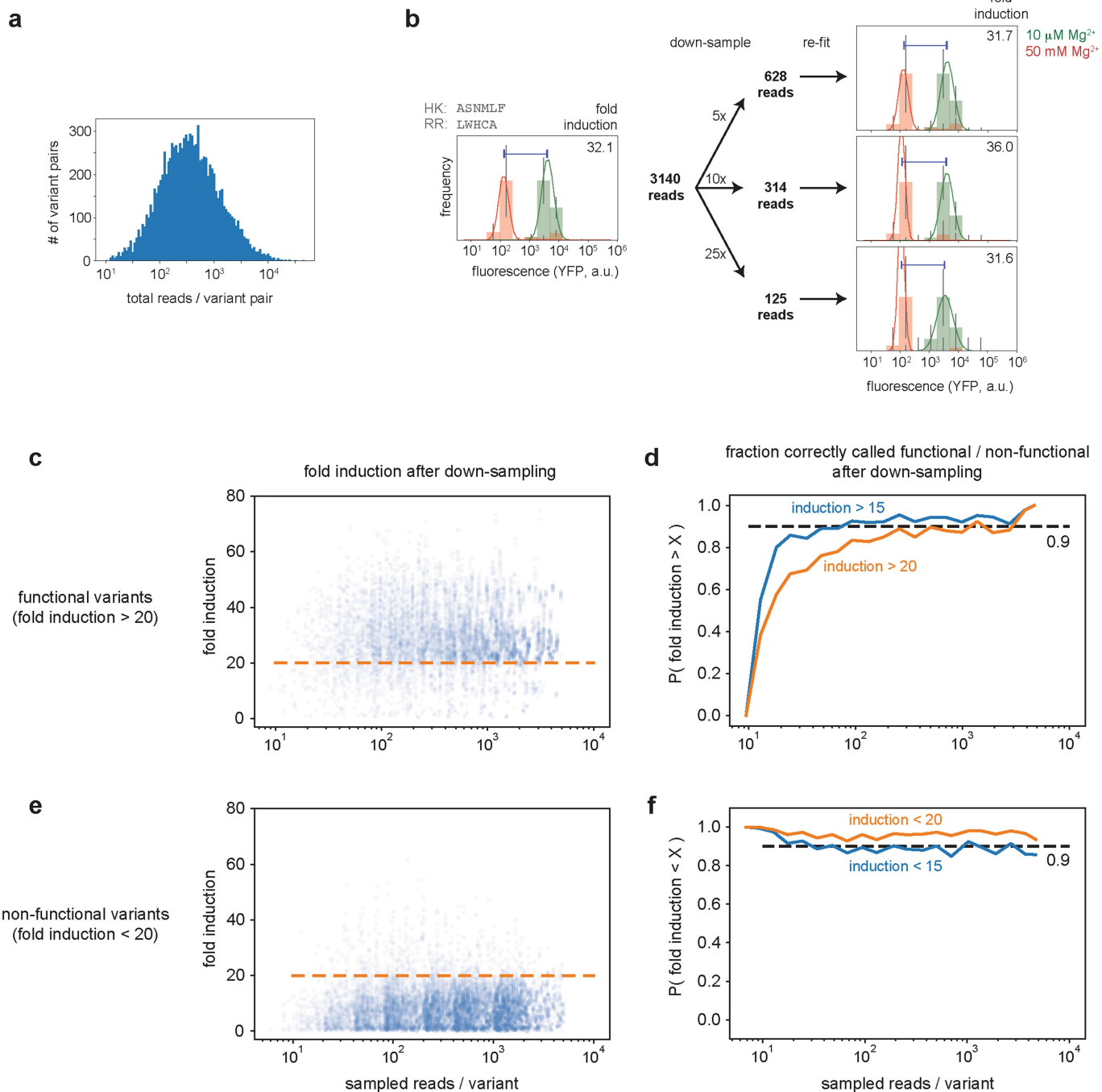
incompatible with all of the paralogues that are already present. **b**, Schematic summarizing the endogenous two-component pathways in *E. coli* to which a new orthogonal pathway must avoid crosstalk. **c**, Diagram of the DHp domain of the histidine kinase TM0853 (blue) in complex with its cognate response regulator RR0468 (green). Residues that dictate specificity, and which were randomized in our libraries, are space-filled in orange (kinase) and red (substrate). **d**, Plot summarizing the number of histidine kinases and response regulators in bacterial genomes. **e**, Visualization of the GREMLIN model, representing the coevolutionary dependencies between the residues of cognate histidine kinases and response regulators. Blue nodes indicate PhoQ residues, green nodes indicate PhoP residues and the darker nodes are the 11 residues that were randomized in the dual PhoQ–PhoP library. Edge widths indicate the strength of the coevolutionary signal, and the node size of each residue represents the total coevolutionary signal to residues on the other protein.



### Extended Data Fig. 2 | Summary statistics for the dual PhoQ–PhoP library.

**a**, Schematic summary of library design. **b**, Left, histogram of the read counts for the pre-selection PhoQ–PhoP library. The vast majority of reads are unique, which indicates that the library size is larger than Illumina sequencing coverage and no variants are overrepresented. Right, histogram of the read counts for one replicate of the PhoQ–PhoP library after overnight growth in low  $Mg^{2+}$  conditions. **c**, Counts of cells sorted into each bin, for each replicate and growth condition. **d**, The cells sorted into each bin were grown overnight, diluted back to mid-exponential phase, shifted to medium with  $10 \mu M Mg^{2+}$  and their YFP levels were verified by flow cytometry.  $n = 2$  independent biological replicates. **e**, As in **d**, but with cells retained in medium with  $50 mM Mg^{2+}$ .  $n = 2$  independent biological replicates. **f**, Scatter plots displaying the correlations between the bin

frequencies of individual variant pairs measured in independent replicates. Only  $10^6$  data points are shown for clarity.  $R^2$  values indicate the Pearson correlation coefficients, calculated using all data points. **g**, As in **f**, but displaying only the 10,595 variants with sufficient sequencing coverage and fit quality (Methods) to be included in the analysis. **h**, Scatter plot displaying the Pearson correlation between the YFP fold induction measured by sort-seq and that measured individually by flow cytometry for 32 individual variant pairs. **i**, Sequence logos summarizing the amino acid frequencies at each position varied in the pre-selected library (top), set of pairs with >20-fold induction (middle) and all native histidine kinase–response regulator pairs (bottom). The residues found at these positions in wild-type PhoQ and PhoP are listed below. **j**, The FACS gating strategy for isolating single live cells for quantification of YFP expression.

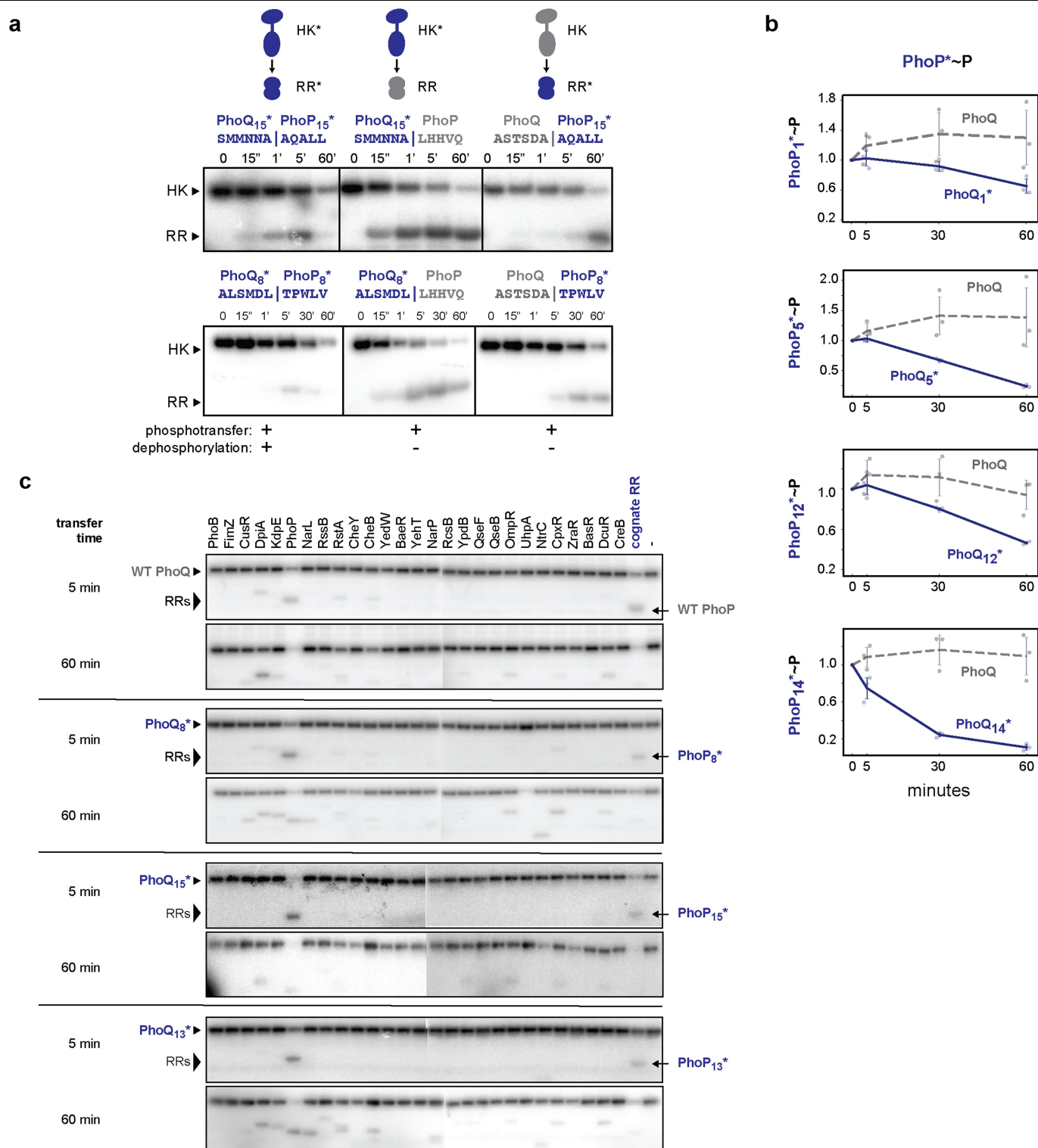


**Extended Data Fig. 3 | Sensitivity of sort-seq pipeline to read coverage.**

**a**, Histogram of the total read counts for the 10,595 variants with sufficient coverage and fit quality to be included in analysis. **b**, A schematic example of the downsampling method used to simulate low read coverage using variants with high read coverage. **c**, Fold induction of high-coverage functional PhoQ<sup>\*</sup>-PhoP<sup>\*</sup> variants (fold induction > 20) after simulating lower read coverage using downsampling and refitting.  $n = 100$  independent downsampling simulations. **d**, A quantification of how read coverage in **c** affects the calculated fold

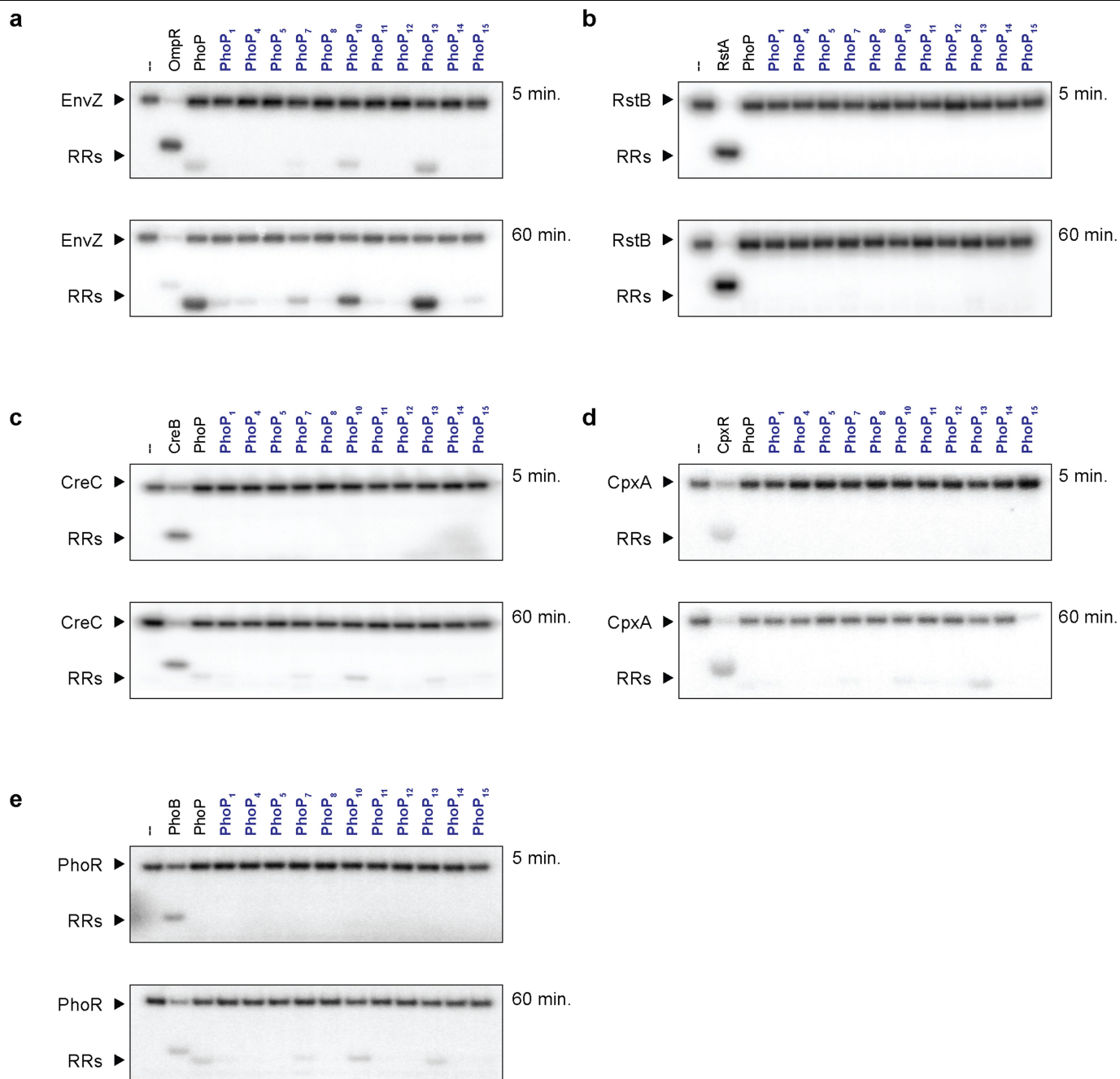
induction of functional variants. The fraction of functional variants that continue to display high fold induction at lower read coverage is plotted with respect to read coverage. **e**, As in **c**, but for nonfunctional (fold induction < 20) PhoQ<sup>\*</sup>-PhoP<sup>\*</sup> pairs with high read coverage.  $n = 100$  independent downsampling simulations. **f**, As in **d**, but for nonfunctional variants. The fraction of nonfunctional variants that continue to display low fold induction at lower read coverage is plotted with respect to read coverage.





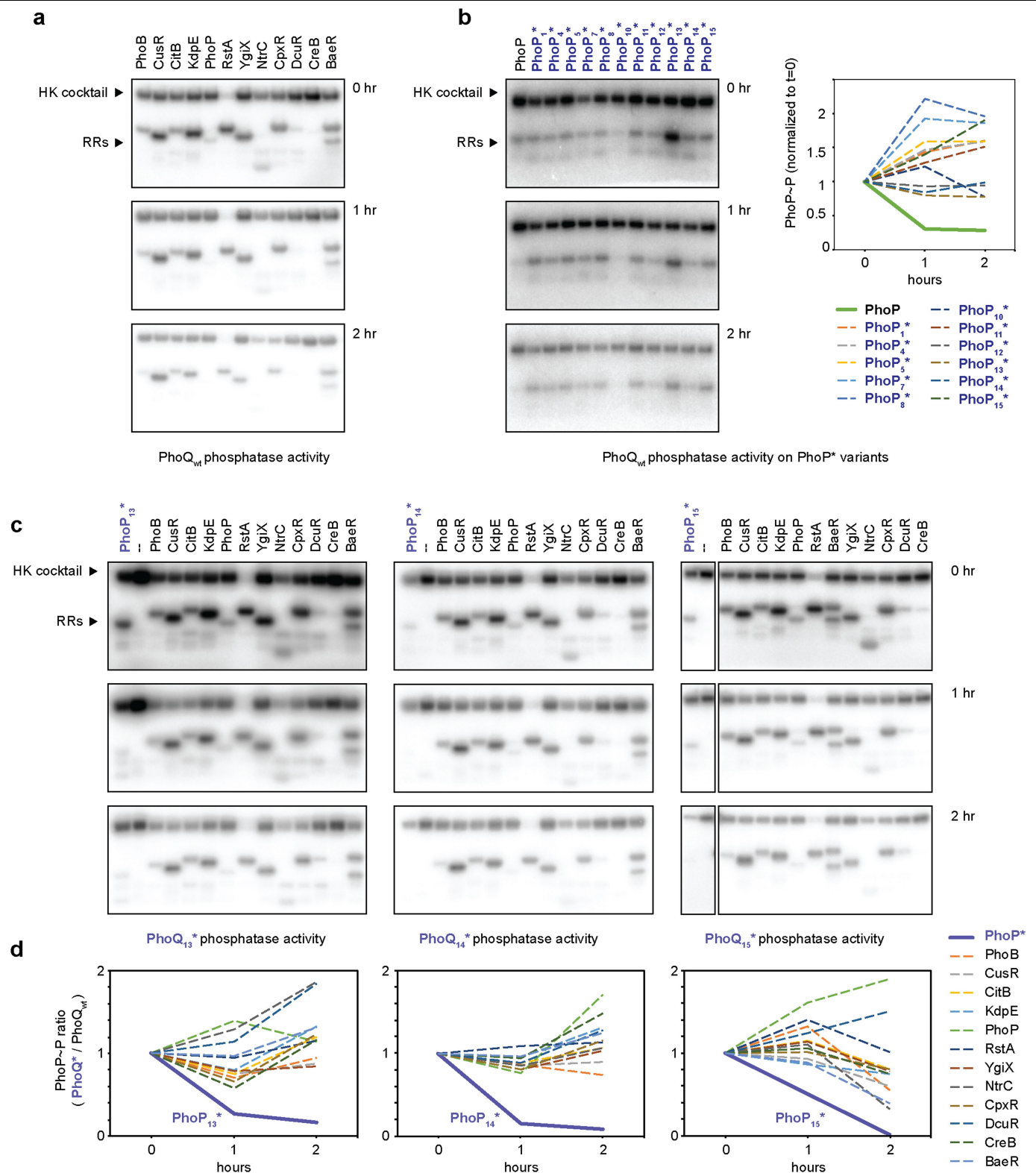
**Extended Data Fig. 4 | Additional in vitro characterizations of selected PhoQ\*–PhoP\* pairs.** **a**, Phosphotransfer reactions for PhoQ\*–PhoP\* variants, as well as PhoQ\* with wild-type PhoP (middle) and wild-type PhoQ with PhoP\* (right). These experiments were repeated independently twice with similar results. **b**, Quantification of phosphatase activity for PhoQ<sub>1</sub>\*, PhoQ<sub>5</sub>\*, PhoQ<sub>12</sub>\* and PhoQ<sub>14</sub>\*, as in Fig. 2c. Lines indicate mean  $\pm$  s.d. from  $n = 3$  biological replicates. **c**, Phosphotransfer profiles of wild-type PhoQ (top) and three PhoQ\* variants, as

in Fig. 2d, but for 60 min and 5 min instead of just 5 min. In each case, the kinase was autophosphorylated and then incubated for 5 or 60 min individually with each of 27 response regulators from *E. coli*, and its selected PhoP\* variant, followed by SDS–PAGE and autoradiography. The position of the autophosphorylated kinase and the approximate positions of any phosphorylated regulators are indicated by arrowheads on the left. These experiments were repeated independently twice with similar results.



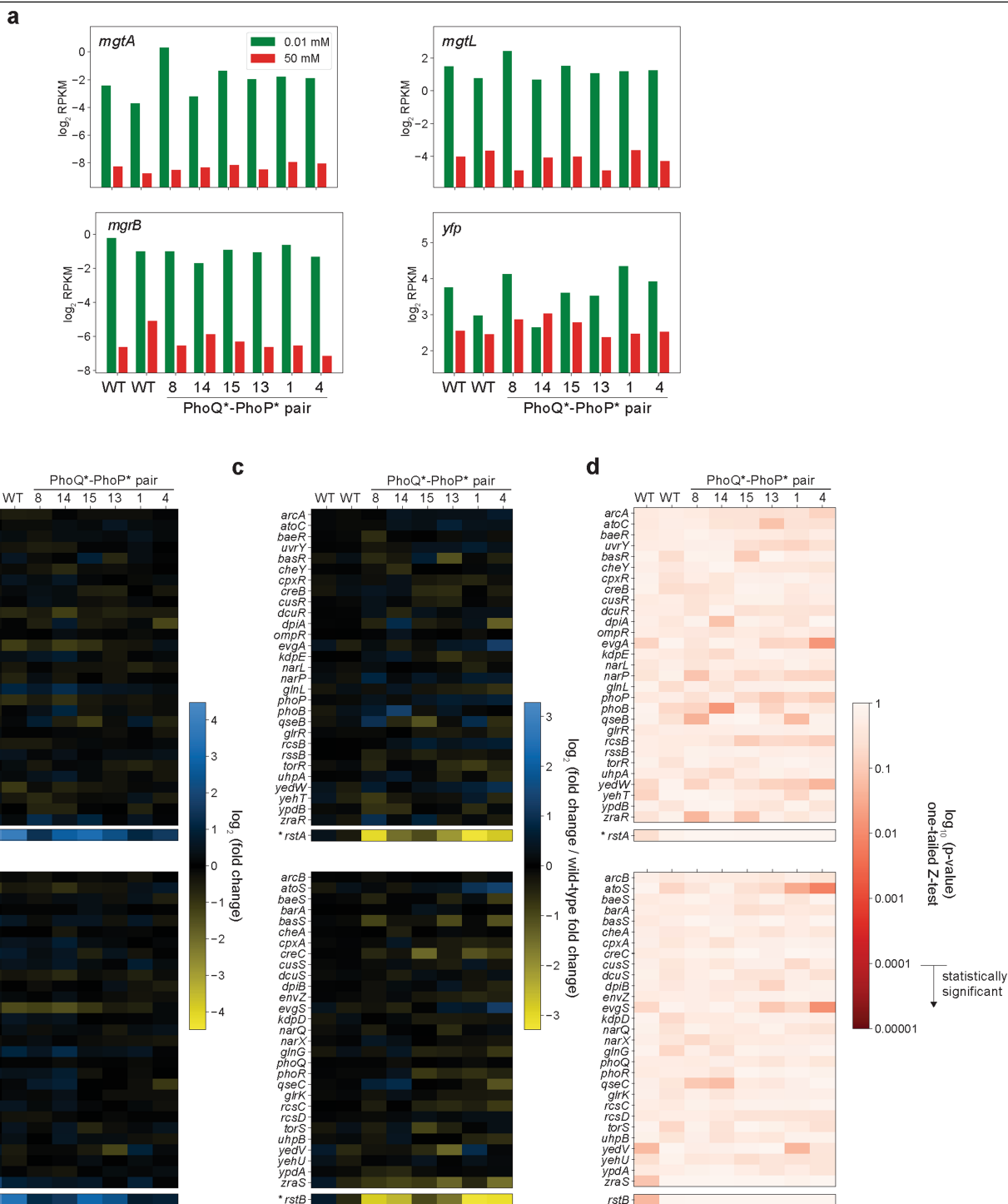
**Extended Data Fig. 5 | Insulation of *E. coli* histidine kinases from PhoP\* variants. a–e.** Phosphotransfer profiles of five histidine kinases endogenous to *E. coli*: EnvZ (a), RstB (b), CreC (c), CpxA (d) and PhoR (e). In each case, the kinase

was autophosphorylated, then incubated for 5 or 60 min. with its cognate response regulator, with wild-type PhoP, or with one of eleven PhoP\* variants, and analysed as in Extended Data Fig. 4.  $n = 1$  independent experiment.



**Extended Data Fig. 6 | Insulation of PhoQ\* variants from *E. coli* response regulators, with respect to phosphatase activity.** **a**, Phosphatase activity of PhoQ was assessed by measuring the decay of phosphorylated response regulators. Twelve *E. coli* response regulators were selected for their ability to be stably phosphorylated in vitro by a cocktail of six *E. coli* histidine kinases (CreC, RstA, PhoR, PhoP, EnvZ and CpxA, each at 250 nM). After 2 h of pre-incubation with radiolabelled ATP and this kinase cocktail, each regulator was combined with 2 mM PhoQ and the phosphorylation state of the regulators was measured

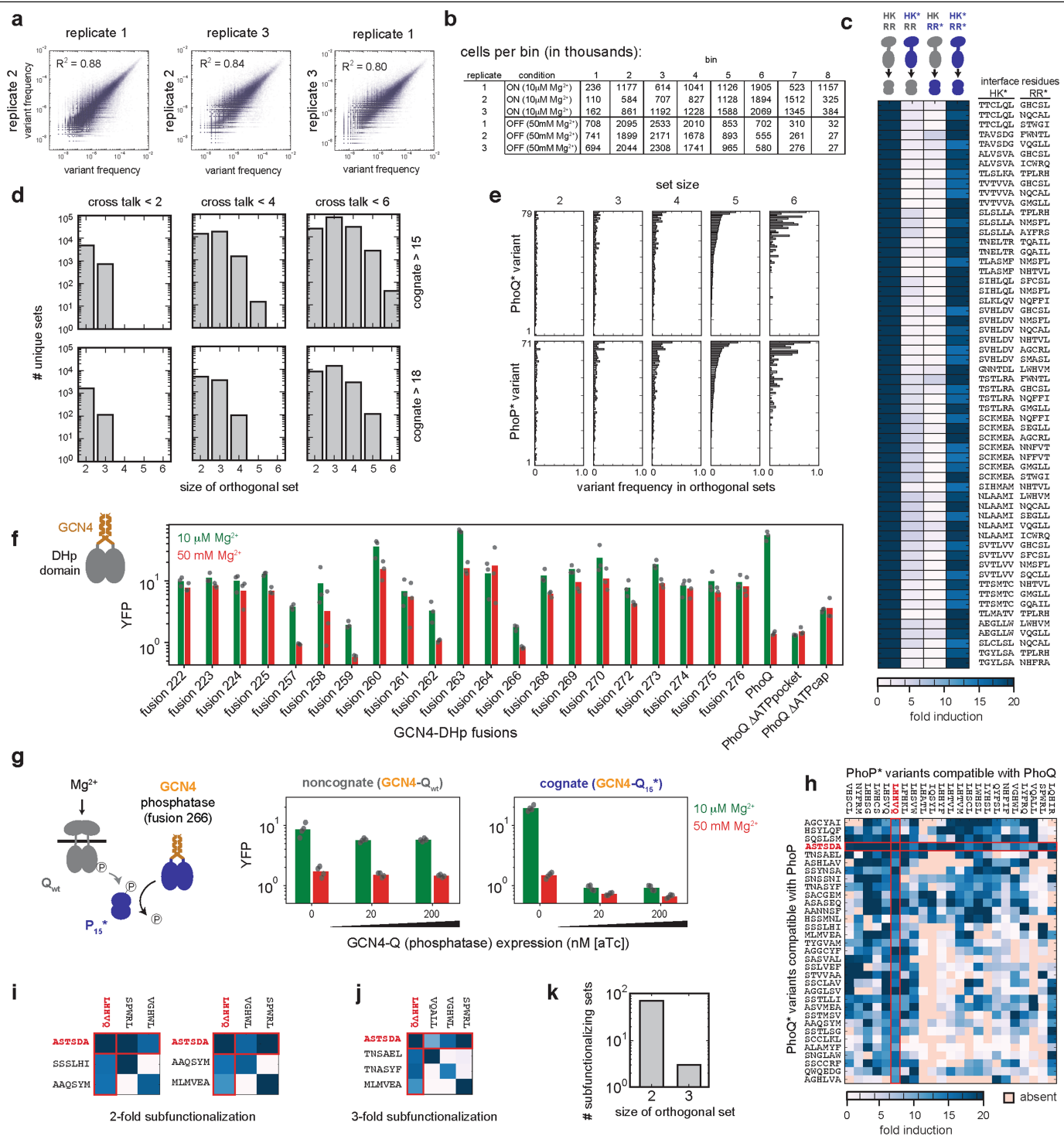
after 0, 60 and 120 min.  $n=1$  independent experiment. **b**, Phosphatase profiles conducted as in **a** for PhoP\* variants. Quantification of wild-type PhoP and PhoP\* variant phosphorylation (normalized to 0 h ( $t=0$ ) to display decay) is plotted on the right.  $n=1$  independent experiment. **c**, Phosphatase profiles conducted as in **a** for PhoQ\* variants.  $n=1$  independent experiment. **d**, Ratio of response regulator phosphorylation between phosphatase profiles with PhoQ\* variants (**c**) and wild-type PhoQ (**a**, **b**).



**Extended Data Fig. 7 | RNA-seq analysis of strains containing PhoQ\*-PhoP\* variants.** **a**, RNA-seq analysis of strains containing wild-type PhoQ-PhoP or the indicated variant pair, measured after 30 min with 10  $\mu$ M or 50 mM  $Mg^{2+}$ . Each strain displays a similar  $Mg^{2+}$ -limitation induction of three genes (*mgtA*, *mgtL* and *mgrB*) in the PhoP regulon, as well as the PhoP-dependent reporter gene *yfp*. **b**, The expression change of each response-regulator and histidine-kinase gene in *E. coli*. Colours represent the fold change (expressed in  $\log_2$ ) in the response to low extracellular  $Mg^{2+}$ . *rstA* and *rstB* are part of the PhoP regulon, and so show changes in transcription after activation of PhoQ or of several PhoQ\* variants. Otherwise, most two-component pathways show little induction by the wild-type PhoQ-PhoP and the PhoQ\*-PhoP\* pairs. **c**, The same data as in **b**, but with the fold change of each variant pair normalized to the fold change seen with wild-

type PhoQ-PhoP, in which the latter is the geometric mean of two wild-type replicates. **d**, *P* values of the z-score calculated for each value in **c**. For each gene and each variant, z-scores represent the deviation of the variant/wild-type ratio of that gene, when compared to the distribution of the variant/wild-type ratio of every gene. Using all *E. coli* genes with reads across multiple samples ( $n = 3,477$ ), *P* values were calculated with a one-tailed z-test to identify genes that induced more strongly with the variant pairs than with the wild-type pair. The statistical significance threshold after correcting for multiple hypothesis testing is indicated on the colour legend that encodes the *P* values. None of the other two-component signalling genes in *E. coli* is significantly induced by the variant PhoQ\*-PhoP\* pairs that we tested.

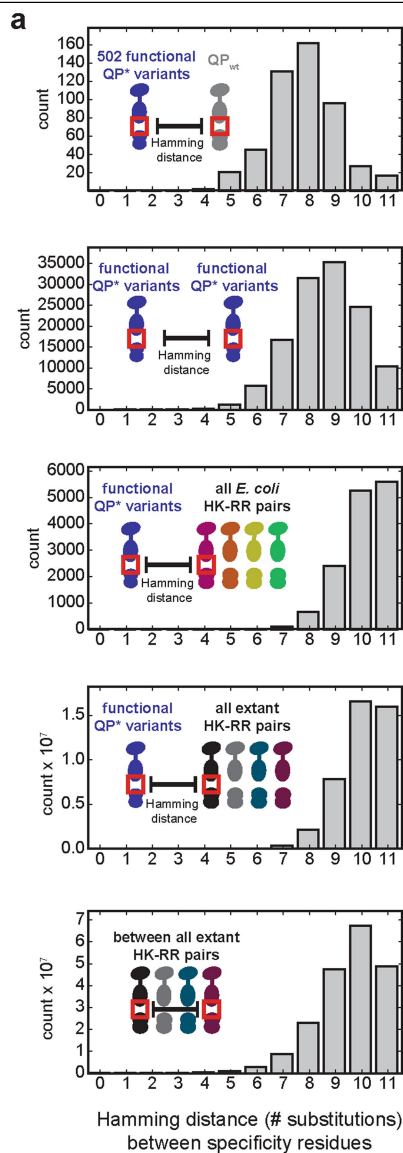




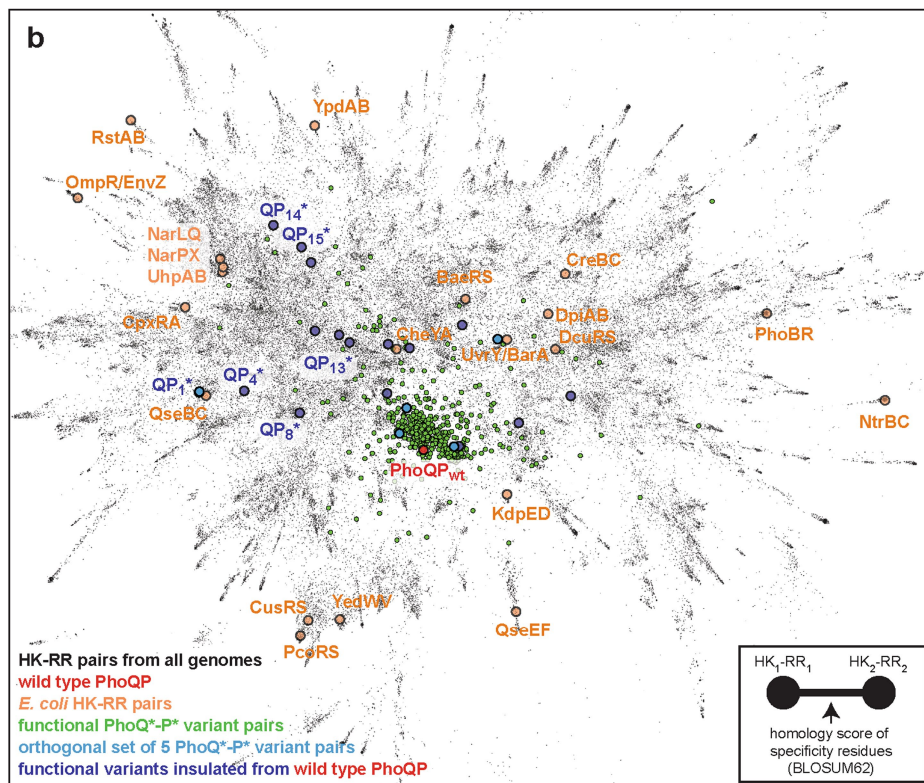
**Extended Data Fig. 8 | Additional characterization of orthogonal sets of PhoQ\*–PhoP\* variant pairs.** **a**, Reproducibility of replicates for the combinatorial library in Fig. 3a. Correlations between bin frequencies of individual variant pairs measured in independent replicates.  $R^2$  values indicate the Pearson correlation coefficients, calculated using all data points ( $n = 210,319$  variant pairs). **b**, Counts of cells sorted into each bin, for each replicate and growth condition. **c**, Functional PhoQ\*–PhoP\* variants that are orthogonal to wild-type PhoQ and PhoP. The fold-induction values, taken from the matrix in Fig. 3a, measured by sort-seq for the variant pairs indicated in each row, either together (far right column of the heat map) or with a wild-type protein (middle two columns), compared to the wild-type pair (far left column). **d**, Number of unique sets of various sizes of orthogonal PhoQ\*–PhoP\* pairs, for various thresholds of activity and crosstalk. **e**, Frequency of each PhoQ\* (top) or

PhoP\* (bottom) variant within the orthogonal sets of various sizes (fold induction > 15, crosstalk < 6). **f**, Phosphatase- and kinase-locked variants of PhoQ were identified by fusing the catalytic DHp–CA domains to the leucine zipper GCN4 at different fusion sites (Methods). **g**, Phosphatase-locked PhoQ<sub>15</sub>\* is sufficient to suppress nonspecific phosphotransfer from wild-type PhoQ to PhoP<sub>15</sub>\*. **h**, Heat map as in Fig. 3a, but restricted to variants that retain an interaction (fold induction > 10) with wild-type PhoQ and PhoP, which are shown in red. **i, j**, Orthogonal sets of PhoQ\* and PhoP\* variants that, similar to the set in Fig. 3h, comprise exclusively proteins that retain interactions with the parent PhoQ and PhoP. **k**, Number of unique sets of various sizes of orthogonal PhoQ\*–PhoP\* pairs in which all variants retain an interaction (fold induction > 10) with wild-type PhoP and PhoQ.

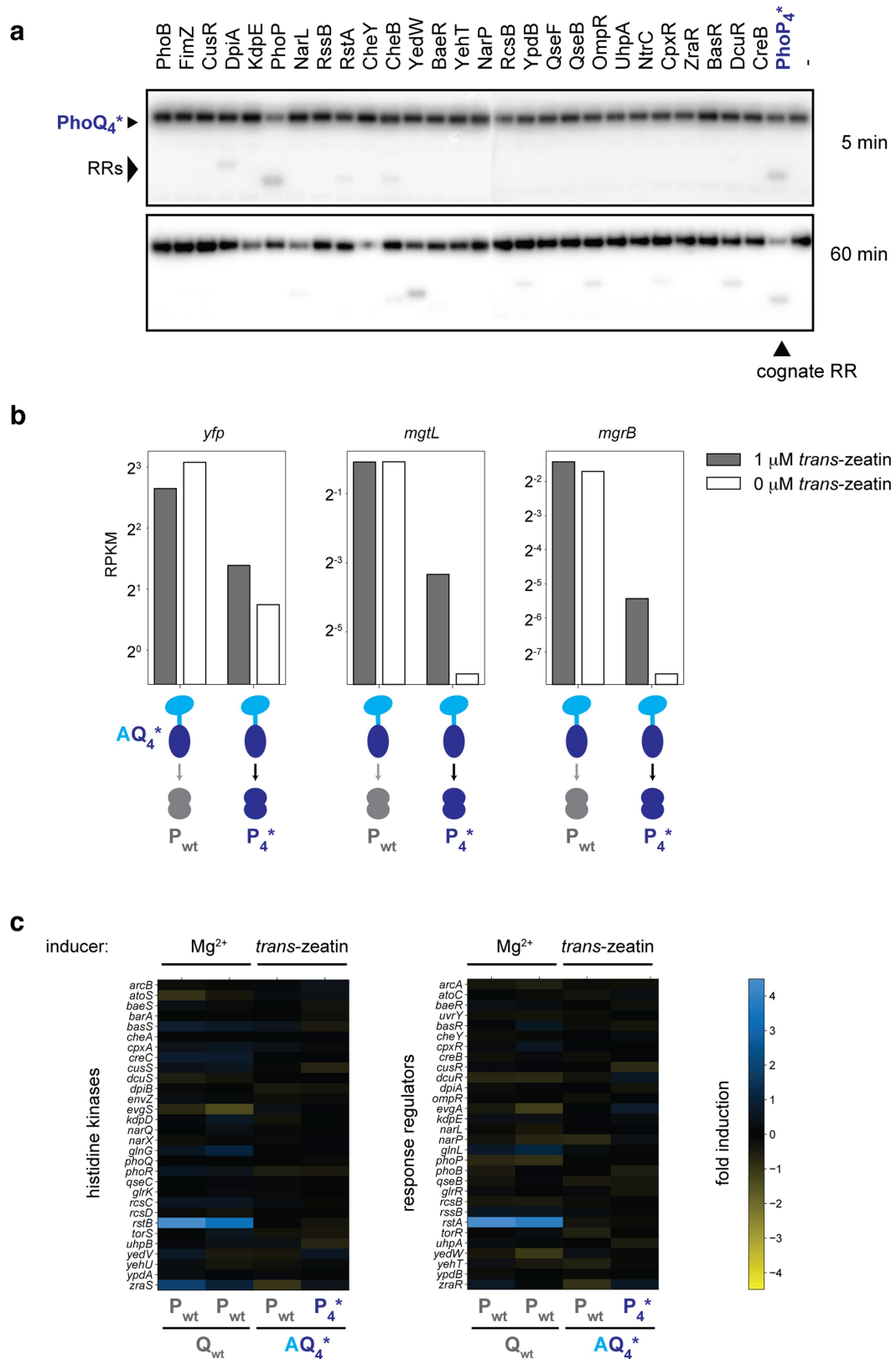




**Extended Data Fig. 9 | Specificity residues of novel PhoQ\*-PhoP\* pathways are distinct from all extant two-component signalling interfaces. a,** Hamming distance was calculated between the 11 specificity residues of each PhoQ\*-PhoP\* variant pair, and the 11 specificity residues of wild-type PhoQ-PhoP, the two-component signalling paralogs in *E. coli* or all extant two-component signalling proteins. When comparing two sets, all distances



between all members of both sets are plotted. **b,** Force-directed graph representing the sequence space of histidine kinases. Each node represents a single histidine kinase, with the relative positions reflecting the similarity of their interface residues (Methods). Coloured nodes highlight specific sets of kinases, as indicated in the legend.



**Extended Data Fig. 10 | Global insulation of the AQ<sub>4</sub>\*-PhoP<sub>4</sub>\* chimeric signalling pathway. a**, Phosphotransfer profile of PhoQ<sub>4</sub>\*. PhoQ<sub>4</sub>\* was autophosphorylated and then incubated for 5 and 60 min with each of 27 response regulators from *E. coli* and with PhoP<sub>4</sub>\*, as in Extended Data Fig. 4c. This experiment was repeated independently twice with similar results. **b, c**, RNA-seq analysis, as in Extended Data Fig. 7b, c, of strains that express AQ<sub>4</sub>\*

and either wild-type PhoP or PhoP<sub>4</sub>\* (measured after 30 min induction with 0 mM or 1 mM *trans*-zeatin). **b**, The PhoP-regulated genes *yfp*, *mgtL* and *mgrB* are induced by *trans*-zeatin only when AQ<sub>4</sub>\* is paired with PhoP<sub>4</sub>\*. **c**, The expression change of all response regulators and histidine kinases (fold change in response to 10 mM Mg<sup>2+</sup> or 1 mM *trans*-zeatin).

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated
- ☐ ☒ Clearly defined error bars  
*State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on [statistics for biologists](#) may be useful.

### Software and code

Policy information about [availability of computer code](#)

#### Data collection

Gremlin (openseq.org) was used to identify networks of coevolving residues within the two-component protein complex. The RefSeq Prokaryotic Genomes database of 5,506 bacterial genomes (Sept, 2017) was downloaded from NCBI. The HMMER3 package was used to identify and align response regulator and histidine kinase sequences. Gephi v0.9 and Force Atlas 2 were used to generate the force directed graph. Flow cytometry data was analyzed in FlowJo 10.

#### Data analysis

Illumina reads for Sort-seq experiments were analyzed with custom Python 2.7 code found at <https://github.com/mcclune/nature2019>. RNA-seq reads were mapped to the E. coli genome and plasmids with Bowtie using default parameters. Flow cytometry data was analyzed with FlowJo 10.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Datasets generated during this study have been deposited in GEO. Raw reads and processed Sort-Seq analysis of each mutant can be found under the accession numbers GSE120780 (degenerate PhoQ-PhoP library) and GSE120786 (combinatorial library of 79 PhoQ\* and 71 PhoP\* variants). Raw reads and RPKM for all *Echerichia coli* genes from RNA-seq are deposited with the accession number GSE128611.

## Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://nature.com/authors/policies/ReportingSummary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation is performed; sampling was constrained by limitations of cell sorting and Illumina sequencing depth, but only variants with sufficient coverage (see Data Exclusions section) were included in analysis and conclusions. We performed the Sort-Seq experiments in three biological replicates to assess the reproducibility of these measurements. The means and standard deviations between these replicates were used to fit Gaussian distributions to each variant in each condition.
Data exclusions	Variants with fewer than 25 total reads were discarded before fitting. Poor Gaussian fits have high variances on the estimated parameters. The standard deviation error on the estimated log(YFP) mean (ofit) was used as a metric to filter poorly fit sequences: sequences were removed if ofit, ON + ofit, OFF > 2. Exclusion criteria were not pre-established but were chosen using variants with STOP codons, which served as negative controls.
Replication	All in vivo experiments, including library selections and sorting, were successfully replicated in triplicate.
Randomization	This study does not involve subjects that require randomization.
Blinding	This study does not involve procedures that require blinding.

## Reporting for specific materials, systems and methods

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Flow Cytometry

### Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

Sample preparation	5µL cells were added to 195 µL PBS with 0.5 g/L kanamycin to halt YFP production.
Instrument	Miltenyi MACSQuant VYB
Software	FlowJo 10
Cell population abundance	Cytometry was used to take YFP measurements from entire E. coli cultures. YFP was measured from 100% of the cells in each culture.
Gating strategy	Cells were isolated from cell debris by gating on SSC and FSC. Single cells were isolated by gating on side scatter height and side scatter area. Live cells were isolated by gating on CFP. (See Fig. S2j)

- ☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.



# Recurrent noncoding U1 snRNA mutations drive cryptic splicing in SHH medulloblastoma

<https://doi.org/10.1038/s41586-019-1650-0>

Received: 13 August 2018

Accepted: 3 September 2019

Published online: 9 October 2019

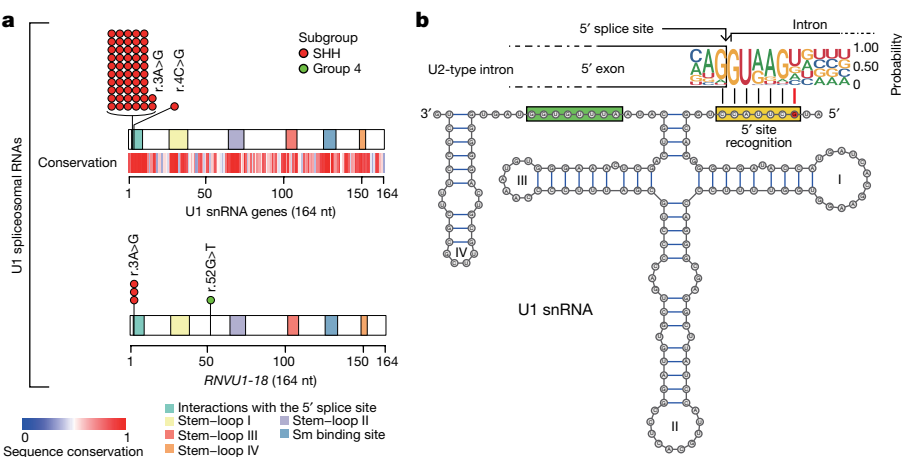
Hiromichi Suzuki<sup>1,2,65</sup>, Sachin A. Kumar<sup>1,2,3,65</sup>, Shimin Shuai<sup>4,5</sup>, Ander Diaz-Navarro<sup>6,7</sup>, Ana Gutierrez-Fernandez<sup>6,7</sup>, Pasqualino De Antonellis<sup>1,2</sup>, Florence M. G. Cavalli<sup>1,2</sup>, Kyle Juraschka<sup>1,2,3</sup>, Hamza Farooq<sup>1,2,3</sup>, Ichiyo Shibahara<sup>1,2</sup>, Maria C. Vladoiu<sup>1,2,3</sup>, Jiao Zhang<sup>1,2</sup>, Namal Abeysundara<sup>1,2</sup>, David Przelicki<sup>1,2,3</sup>, Patryk Skowron<sup>1,2,3</sup>, Nicole Gauer<sup>1,2</sup>, Betty Luu<sup>1,2</sup>, Craig Daniels<sup>1,2</sup>, Xiaochong Wu<sup>1,2</sup>, Antoine Forget<sup>8,9</sup>, Ali Momin<sup>1,2,5</sup>, Jun Wang<sup>10</sup>, Weifan Dong<sup>1,2,5</sup>, Seung-Ki Kim<sup>11</sup>, Wiesława A. Grajkowska<sup>12</sup>, Anne Jouvett<sup>13</sup>, Michelle Fèvre-Montange<sup>14</sup>, Maria Luisa Garré<sup>15</sup>, Amulya A. Nageswara Rao<sup>16</sup>, Caterina Giannini<sup>17</sup>, Johan M. Kros<sup>18</sup>, Pim J. French<sup>19</sup>, Nada Jabado<sup>20</sup>, Ho-Keung Ng<sup>21</sup>, Wai Sang Poon<sup>22</sup>, Charles G. Eberhart<sup>23,24,25</sup>, Ian F. Pollack<sup>26</sup>, James M. Olson<sup>27</sup>, William A. Weiss<sup>28,29,30</sup>, Toshihiro Kumabe<sup>31</sup>, Enrique López-Aguilar<sup>32</sup>, Bolesław Lach<sup>33,34</sup>, Maura Massimino<sup>35</sup>, Erwin G. Van Meir<sup>36,37,38</sup>, Joshua B. Rubin<sup>39,40</sup>, Rajeev Vibhakkar<sup>41</sup>, Lola B. Chambless<sup>42</sup>, Noriyuki Kijima<sup>43</sup>, Almos Klekner<sup>44</sup>, László Bognár<sup>44</sup>, Jennifer A. Chan<sup>45</sup>, Claudia C. Faria<sup>46,47</sup>, Jiannis Ragoussis<sup>48,49</sup>, Stefan M. Pfister<sup>50,51,52</sup>, Anna Goldenberg<sup>53,54</sup>, Robert J. Wechsler-Reya<sup>10,55</sup>, Swneke D. Bailey<sup>56,57</sup>, Livia Garzia<sup>57,58</sup>, A. Sorana Morrissey<sup>45,59</sup>, Marco A. Marra<sup>60,61</sup>, Xi Huang<sup>1,2</sup>, David Malkin<sup>62</sup>, Olivier Ayrault<sup>8,9</sup>, Vijay Ramaswamy<sup>2,62</sup>, Xose S. Puente<sup>6,7</sup>, John A. Calarco<sup>63</sup>, Lincoln Stein<sup>4</sup> & Michael D. Taylor<sup>1,2,3,64\*</sup>

In cancer, recurrent somatic single-nucleotide variants—which are rare in most paediatric cancers—are confined largely to protein-coding genes<sup>1–3</sup>. Here we report highly recurrent hotspot mutations (r.3A>G) of U1 spliceosomal small nuclear RNAs (snRNAs) in about 50% of Sonic hedgehog (SHH) medulloblastomas. These mutations were not present across other subgroups of medulloblastoma, and we identified these hotspot mutations in U1 snRNA in only <0.1% of 2,442 cancers, across 36 other tumour types. The mutations occur in 97% of adults (subtype SHHδ) and 25% of adolescents (subtype SHHα) with SHH medulloblastoma, but are largely absent from SHH medulloblastoma in infants. The U1 snRNA mutations occur in the 5′ splice-site binding region, and snRNA-mutant tumours have significantly disrupted RNA splicing and an excess of 5′ cryptic splicing events. Alternative splicing mediated by mutant U1 snRNA inactivates tumour-suppressor genes (*PTCH1*) and activates oncogenes (*GLI2* and *CCND2*), and represents a target for therapy. These U1 snRNA mutations provide an example of highly recurrent and tissue-specific mutations of a non-protein-coding gene in cancer.

Medulloblastoma, a cerebellar neuronal cancer, comprises four molecular subgroups (WNT, SHH, group 3 and group 4), each of which has its own distinct clinical, transcriptomic and genetic make-up<sup>4–6</sup>. Each of these four molecular subgroups can be further subdivided into subtypes; for SHH medulloblastoma, these comprise SHHα, SHHβ, SHHγ and SHHδ<sup>7</sup>. Noncoding single-nucleotide variants have recently been discovered in the promoter regions of *TERT* and a handful of other loci, which has given impetus to the careful examination of noncoding segments<sup>8,9</sup>. Thus, we sought to explore the genomic landscape of medulloblastoma with a particular focus on noncoding regions. We analysed whole-genome sequencing data of 114 medulloblastomas, and observed a recurrent hotspot mutation of noncoding U1 snRNA genes in 10 out of 114 cases (8.8%) (Fig. 1a, Extended Data Fig. 1, Supplementary Tables 1, 2, Methods). Hotspot mutations of U1 snRNA genes occur in the third

nucleotide (r.3A>G), and are restricted to SHH medulloblastoma. These hotspot mutations are localized within the 5′ splice-site recognition sequence, which has been ultra-conserved in eukaryotes through nearly one billion years of evolution (Fig. 1b, Extended Data Fig. 2a). The human reference genome (hg19) has four annotated U1 snRNA genes (*RNU1-1*, *RNU1-2*, *RNU1-3* and *RNU1-4*) and three ‘pseudogenes’ (*RNU1-27P*, *RNU1-28P* and *RNVU1-18*), all of which encode completely identical 164-base-pair transcripts. In addition, there are over 100 U1 snRNA pseudogenes spread across the genome, which highly complicates their identification by mutation callers owing to the inability to align short reads to any individual U1 snRNA gene<sup>10</sup> (Extended Data Fig. 3). We remapped sequence reads that permitted multimapping, and successfully detected the U1 snRNA mutation in five additional cases (Methods). We validated hotspot U1 snRNA mutations in an additional 40 out of 227 cases of

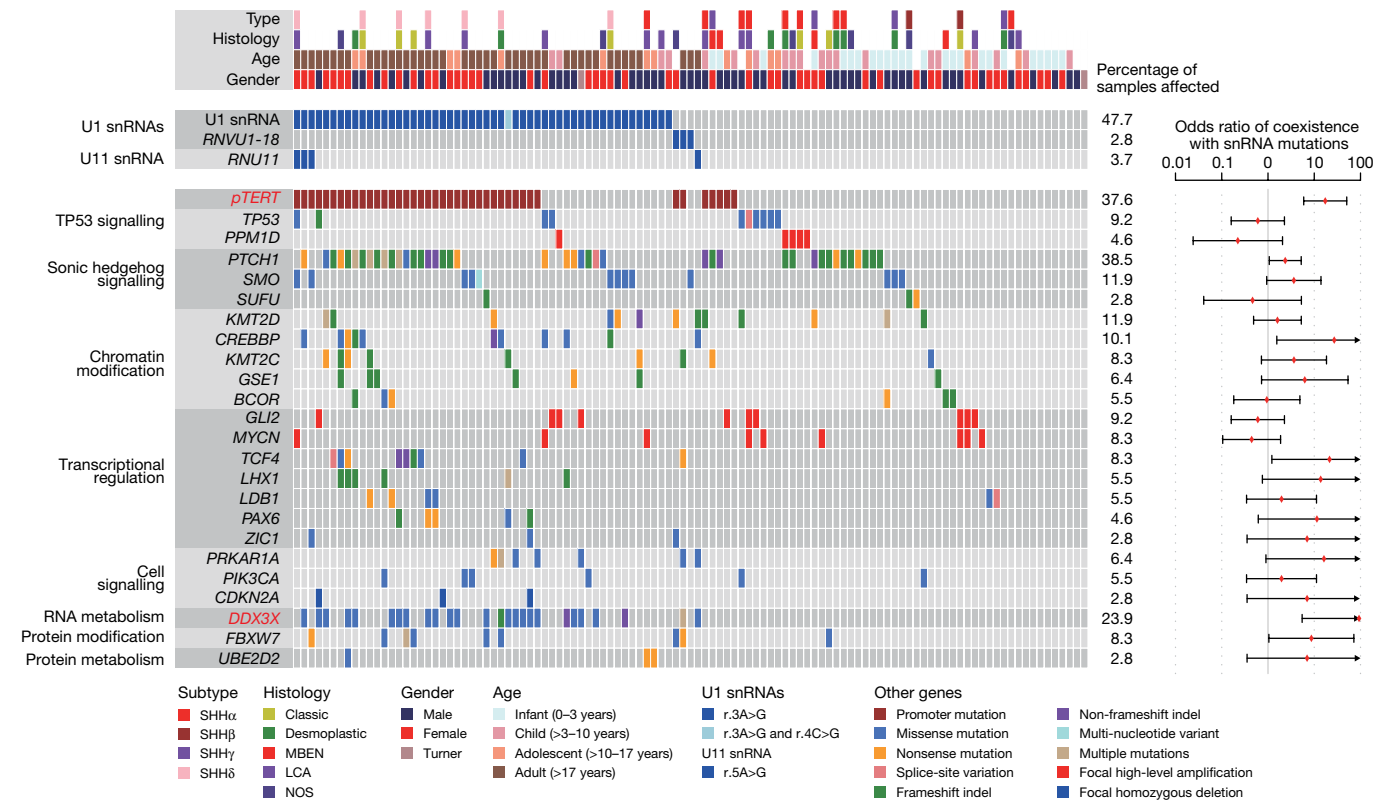
A list of affiliations appears at the end of the paper.



**Fig. 1 | Highly recurrent mutations of the U1 snRNAs in SHH medulloblastoma.** **a**, Cartoon illustrating the number and subgroup-specific distribution of somatic mutations in the U1 snRNA genes. U1 snRNA sequence conservation scores, as determined using the Rfam database (<http://rfam.xfam.org/>). **b**, Secondary structure of the mutant U1 snRNA. The red circle identifies the location of the hotspot mutation. The yellow and green rectangles indicate the 5' splice-site recognition site and the Sm protein-binding site, respectively. Numerals I to IV indicate stem loops.

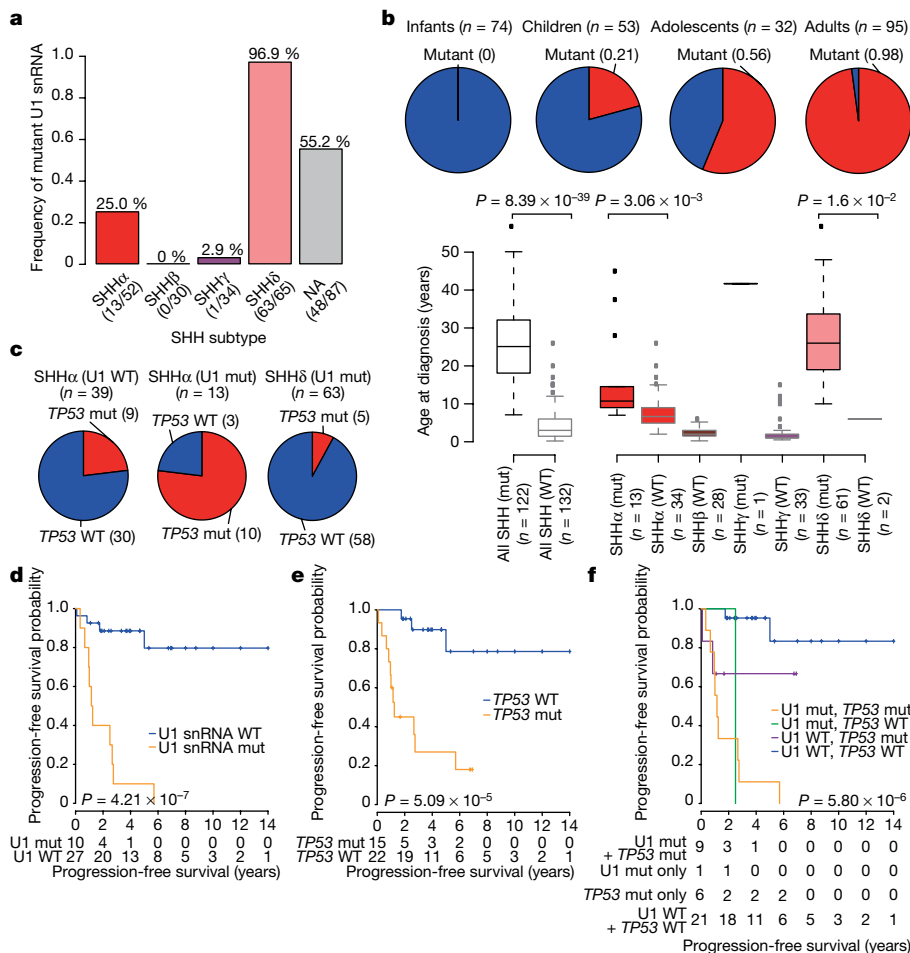
medulloblastoma from the International Cancer Genome Consortium (ICGC) (Supplementary Tables 2–4). We also detected recurrent hotspot mutations of the U1 snRNA gene (*RNU11*) at the fifth nucleotide (r.5A>G), in the highly conserved 5' splice-site recognition sequence (in a total of 4 out of 341 cases) (Extended Data Fig. 2b–d, Supplementary Table 2). Taken together, 51% (56 out of 109) of SHH medulloblastomas have at least one mutation in U1 or U11 snRNA (Fig. 2). These snRNA mutations significantly co-occur with mutations of the *TERT* promoter and *DDX3X* (Supplementary Tables 5, 6). We assessed the U1 snRNA r.3A>G mutations across 2,442 samples drawn from 36 cancer histologies of the ICGC, and found such a mutation in only one sample (0.04%)—a lone sample of pancreatic ductal adenocarcinoma (Supplementary Table 7). We conclude that U1 snRNA r.3A>G mutations are both highly recurrent in, and extremely specific to, SHH medulloblastoma.

We validated the U1 snRNA r.3A>G mutations in an additional 159 cases of SHH medulloblastoma using allele-specific PCR. We detected mutations in the *RNU1-27P* and/or *RNU1-28P* genes (which were confirmed by Sanger sequencing) that were not identified by whole-genome sequencing (Extended Data Fig. 4a, b, Supplementary Table 8, Methods). Combining the results of whole-genome sequencing and allele-specific PCR, we found that U1 snRNA r.3A>G mutations were largely restricted to cases of SHH medulloblastoma in adulthood (SHH $\delta$ , present in 97% of cases) and adolescence (SHH $\alpha$ , present in 25% of cases), and absent from those in infancy (Fig. 3a, b). This remains true if accounting only for age, and not for molecular subtype. Indeed, most patients with SHH $\alpha$  with *TP53* mutations also have U1 snRNA r.3A>G mutations (Fig. 3c). Both broad and focal somatic copy-number variations are divergent between SHH $\alpha$  with wild-type U1 snRNA, SHH $\alpha$  with mutant U1 snRNA



**Fig. 2 | Mutational repertoire of snRNA-mutant SHH medulloblastomas.** Genomic landscape of mutations in SHH medulloblastomas ( $n = 109$ ), with and without U1 or U11 mutations. Odds ratios (red dots) of coexistence of U1 and U11 snRNA mutations with other somatic events are shown with their 95%

confidence intervals. Arrowheads represent values that are out of the axis range. Significantly correlated mutations are denoted in red (false discovery rate (FDR) < 0.1, asymptotic  $P$  values from odds-ratio tests ( $H_0$ : odds ratio = 1, Methods) with Benjamini–Hochberg adjustment for multiple testing).



**Fig. 3 | Clinical and cytogenetic features of SHH medulloblastomas with mutant U1 snRNA.**

**a**, Frequency of U1 snRNA mutations across subtypes of SHH medulloblastoma. NA, not available (samples for which the subtype is unknown). **b**, Top, frequency of U1 snRNA mutation by age group ( $n = 74$  for infants,  $n = 53$  for children,  $n = 32$  for adolescents,  $n = 95$  for adults). Bottom, age distribution by subtype ( $n = 47$  for SHHα,  $n = 28$  for SHHβ,  $n = 34$  for SHHγ,  $n = 63$  for SHHδ,  $n = 82$  for unknown subtype) and U1 snRNA mutational status ( $n = 122$  for mutant,  $n = 132$  for wild type (WT)). P-values were calculated by two-sided Wilcoxon rank-sum test. In the box plots, centre lines show data median; box limits indicate the interquartile range (IQR) from the 25th to the 75th percentiles; and lower and upper whiskers extend  $1.5 \times$  the IQR. Outliers are represented by individual points. **c**, Frequency of TP53 mutation in tumours with wild-type and mutant U1 snRNA. **d–f**, Progression-free survival of patients with SHHα, stratified by mutational status of U1 snRNA (**d**) ( $n = 10$  for mutant,  $n = 27$  for wild type), TP53 (**e**) ( $n = 15$  for mutant,  $n = 22$  for wild type) or both (**f**) ( $n = 9$  for both mutant,  $n = 1$  for U1 snRNA mutation only,  $n = 6$  for TP53 mutation only,  $n = 21$  for both wild type). P-values were determined using the two-sided log-rank test. +, censored cases.

and SHHδ with mutant U1 snRNA, which provides support for a model in which these cancers follow different genetic pathways to transformation (Extended Data Fig. 4c, d, Supplementary Tables 9, 10). An analysis of focal copy-number variations demonstrates that SHHα tumours with wild-type U1 snRNA have an increased incidence of copy-number variations that encompass several oncogenes and tumour-suppressor genes, including *MYCN*, *CCND2* and *PPMID*.

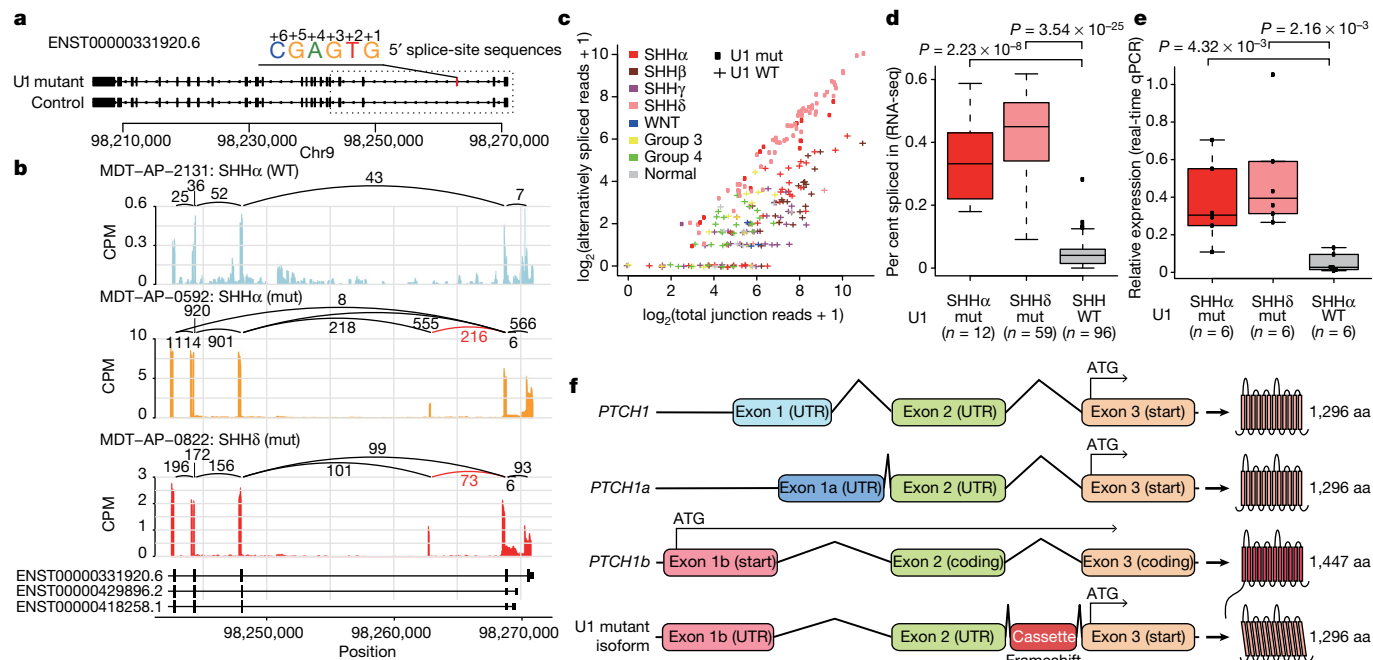
A univariate log-rank analysis of both progression-free survival and overall survival revealed that, within SHHα tumours, both U1 snRNA r.3A>G and TP53 mutational status are separately associated with a significantly poor outcome (Fig. 3d–f, Extended Data Fig. 4e–i). However, in a multivariate Cox regression analysis, TP53 mutations alone are no longer significant for progression-free survival, whereas U1 snRNA r.3A>G mutations confer a very strong risk of relapse (U1 snRNA r.3A>G mutation: hazard ratio 5.51, 95% confidence interval 1.15–26.35,  $P = 0.03$ ; TP53 mutation: hazard ratio 3.01, 95% confidence interval 0.55–16.65,  $P = 0.21$ ). A similar trend was observed for overall survival (U1 snRNA r.3A>G mutation: hazard ratio 3.72, 95% confidence interval 0.74–18.87,  $P = 0.11$ ; TP53 mutation: hazard ratio 2.70, 95% confidence interval 0.46–15.88,  $P = 0.27$ ). This suggests that, within SHHα, the combination of both a TP53 mutation and a U1 snRNA r.3A>G mutation is associated with an extremely poor prognosis.

Intron-centric alternative splicing analysis using LeafCutter confirms that mutant U1 snRNA variants of both SHHα and SHHδ have 2.5–3 times more alternative 5' cryptic splicing events than do SHH medulloblastomas with wild-type U1 snRNA<sup>11</sup> (Extended Data Figs. 5a, b, 6a–c, Supplementary Table 11). The U1 snRNA r.3A>G mutations would be predicted to affect the recognition of the sixth intronic nucleotide from the 5' splice site; indeed, cryptic 5' splice sites recognized in SHH medulloblastoma with mutant U1 snRNA demonstrate enrichment of

a dominant C base, as opposed to the T base observed in tumours with wild-type U1 snRNA (Extended Data Figs. 5c, 6d, e). Pathway analysis of differentially expressed transcripts between SHH medulloblastoma with wild-type versus mutant U1 snRNA demonstrates an increase in nonsense-mediated decay, which is consistent with the destruction of aberrantly spliced transcripts (Extended Data Fig. 7a). To validate the effect of U1 snRNA mutations, we transfected wild-type or mutant U1 snRNA r.3A>G vectors into human embryonic kidney 293T cells, and examined the effects on splicing. Intron-centric analysis clearly demonstrates an enrichment of a C base at the sixth intronic position when using the mutant U1 snRNA vector, and a considerable increase in the incidence of cryptic 5' splicing events that do not overlap with those of SHH medulloblastoma with wild-type U1 snRNA (Extended Data Fig. 7b–d, Supplementary Tables 12, 13).

Clustering on the basis of significant alternative splicing events is clearly driven by U1 snRNA mutational status (Extended Data Fig. 7e, Methods), and tumours with mutant U1 snRNA segregated distinctly from those with wild-type U1 snRNA. We conclude that the U1 snRNA r.3A>G mutation has a marked effect on alternative splicing in affected tumours.

As a complementary approach, we conducted exon-centric alternative splicing analysis using rMATS<sup>12</sup>. We observed that SHH medulloblastoma tumours with mutant U1 snRNA have a higher incidence of cassette exons than do wild-type U1 snRNA controls (Extended Data Figs. 8a–c, 9a, b, Supplementary Table 14). Similar to cryptic 5' alternative splicing events, the dominant base at the sixth intronic base is C (Extended Data Figs. 8d, 9c, Supplementary Table 15). In addition, an increase of retained introns is observed in tumours with mutant U1 snRNA. The 5' splice-site sequences of missed splice sites in retained introns do not have a dominant C at the sixth nucleotide, but instead have the canonical T. This latter result suggests a mechanism in which



**Fig. 4 | Aberrant splicing of Hedgehog signalling genes in SHH medulloblastoma with mutant U1 snRNA.** **a**, Overview of cryptic alternative splicing of *PTCH1*, demonstrating the position of a cryptic cassette exon with the 5' splice-site sequence. **b**, Top, Sashimi plots of splicing of *PTCH1* in representative cases. The bar plot shows counts per million reads (CPM). Numbers refer to the number of junctional reads. Bottom, annotated exon tracks with genomic positions marked. Junctional reads specific to U1 snRNA mutants are in red. **c**, Scatter plot comparing detected alternatively spliced read and total junction reads that shared the 3' splice site. Jittering was performed for both values. **d**, Per cent spliced in values by SHHα with mutant U1 snRNA, SHHδ with mutant U1 snRNA and SHH medulloblastoma with wild-type U1 snRNA (all

subtypes of SHH medulloblastoma). *P* values were calculated using two-sided Wilcoxon rank-sum test. **e**, Box plot of fold changes in expression of the alternatively spliced isoform of *PTCH1* compared to the wild-type isoform of *PTCH1* in subsets of SHH medulloblastoma, as determined by real-time qPCR. Data are mean  $\pm$  s.d. *P* values were calculated using two-sided Wilcoxon rank-sum test. **f**, Illustration of canonical isoforms and the cryptic alternative isoform of *PTCH1*. Putative translation start sites are indicated with an arrow. Resulting proteins (and size) are displayed for each isoform. aa, amino acid; UTR, untranslated region. In the box plots (**d**, **e**), the centre lines show data median; box limits indicate the IQR from the 25th to the 75th percentiles; and lower and upper whiskers extend 1.5 $\times$  the IQR. Outliers are represented by individual points.

mutant U1 snRNA r.3A>G not only recognizes alternative 5' splice sites but also inhibits the wild-type U1 snRNA from detecting canonical splice sites, which results in their aberrant splicing. The retained intron event with the highest per cent spliced in, validated by real-time quantitative (q)PCR, occurs in the gene *PAX6*, (which undergoes frequent somatic mutation in SHH medulloblastoma) and a chromatin remodelling gene *TOX4*<sup>13,14</sup> (Extended Data Figs. 8e–h, 9d, Supplementary Table 16). The retained introns in both genes result in a frameshift, which leads to loss of function. These data may support a model in which U1 snRNA r.3A>G mutations impede normal splicing, which leads to intron retention and to an mRNA frameshift.

To detect pathogenic alternative splicing, we identified cryptic 5' splicing events with a C base at the sixth intronic position that are shared by mutant U1 snRNA variants of both SHHα and SHHδ tumours (Extended Data Fig. 9e, Supplementary Table 17, 18). Using both RNA sequencing and real-time qPCR, we detected cryptic splicing events with high effect sizes in both *PTCH1* and *GLI2*; these events were highly specific to U1 snRNA r.3A>G mutation (in both SHHα and SHHδ tumours), as compared to wild-type U1 snRNA controls (Fig. 4a–e). *PTCH1* is known to have at least three different initial exons. Splicing mediated by the U1 snRNA r.3A>G mutant results in the inclusion of a cassette exon between exon 2 and exon 3, which causes a frameshift, and therefore predicted translation from the ATG in exon 3 (Fig. 4f). It has previously been reported that loss of expression of the 1,447-amino-acid isoform of *PTCH1* results in derepression of Hedgehog signalling<sup>15</sup>. Similarly, the U1 snRNA r.3A>G cassette exon in *GLI2* is spliced between exon 4 and exon 5, which results in a putative GLI2 protein that lacks the repressor domain (Extended Data Fig. 10a–f). Physiological GLI2 protein has a repressor domain at its amino terminus, and constructs that lack the

amino terminus are much more potent at activating Hedgehog signalling than the full-length protein<sup>16</sup>.

Alternative splicing of the cell-cycle gene *CCND2*, a known downstream target of SHH signalling that is recurrently amplified in SHH medulloblastoma, is detected in U1 snRNA r.3A>G mutants of SHHδ but not in SHHα<sup>17,18</sup> (Extended Data Fig. 10g–l). Focal amplifications of *CDK6* are highly recurrent in SHHα U1 snRNA r.3A>G mutants, but not in SHHα with wild-type U1 snRNA or SHHδ U1 snRNA r.3A>G mutants, which suggests convergence on dysregulation of the G1–S cell-cycle checkpoint. The *CCND2* alternative isoform is prematurely terminated, which results in N-terminal sequences in which the PEST domain is predicted to be deleted. Deletion of the PEST domain causes resistance to protein degradation and impaired export from the nucleus, which results in *CCND2* accumulating in the nucleus to promote cell-cycle progression<sup>19</sup>. *PAX5*, another known tumour-suppressor gene, is affected by cryptic 5' alternative splicing in U1 snRNA r.3A>G mutants (Extended Data Fig. 10m–q). SHH medulloblastomas with wild-type and mutant U1 snRNA express distinct cryptic isoforms. The cryptic isoform present in SHH medulloblastomas with wild-type U1 snRNA translates the complete DNA-binding domain of *PAX5*. However, the cryptic exon (also known as a poison exon<sup>20,21</sup>) that is present in SHH medulloblastomas with mutant U1 snRNA results in a stop codon before the DNA-binding domain. Mutations of *PAX5* in cancer are typically concentrated in the DNA-binding site<sup>22</sup>. Together, the data relating to the alternative splicing of *PTCH1*, *GLI2*, *CCND2* and *PAX5* support a model in which cryptic alternative splicing mediated by mutant U1 snRNA r.3A>G functions as a driver in subsets of SHH medulloblastoma.

A U1 snRNA r.3A>G mutation is the most common single-nucleotide variant in medulloblastoma. The restriction of these mutations not

just to SHH medulloblastoma but to the SHH $\alpha$  and SHH $\delta$  subtypes suggests a model in which the specific cell of origin, the temporally specific microenvironment or co-occurring mutations (that is, of *TP53*) are necessary for U1 snRNA to contribute to oncogenesis. Although the almost universal occurrence of U1 snRNA mutations in SHH $\delta$  highly supports their role in tumour initiation, proof of the ongoing role of mutant U1 snRNA r.3A>G in tumour maintenance will await its knockdown in a tumour in which it was the initiating genetic event.

Patients with SHH $\alpha$  with a U1 snRNA r.3A>G mutation are an extremely high-risk population, who should be prioritized for the development of targeted therapies. Drugs that are under development directly target the spliceosome, which may show anti-tumour effects in cancers with spliceosomal mutations<sup>23</sup>. Loss of expression of specific genes through cryptic splicing or intron retention could create opportunities for synthetic lethal approaches. Finally, cryptic splicing in SHH medulloblastoma with mutant U1 snRNA leads to a unique form of post-transcriptional hypermutation; this would be predicted to result in the expression of numerous cell-surface neo-epitopes that are never seen in healthy tissues, and which could be targeted using immunotherapies.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1650-0>.

- Pugh, T. J. et al. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature* **488**, 106–110 (2012).
- Jones, D. T. et al. Dissecting the genomic complexity underlying medulloblastoma. *Nature* **488**, 100–105 (2012).
- Northcott, P. A. et al. The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017).
- Northcott, P. A., Korshunov, A., Pfister, S. M. & Taylor, M. D. The clinical implications of medulloblastoma subgroups. *Nat. Rev. Neurol.* **8**, 340–351 (2012).
- Northcott, P. A. et al. Medulloblastoma comprises four distinct molecular variants. *J. Clin. Oncol.* **29**, 1408–1414 (2011).
- Taylor, M. D. et al. Molecular subgroups of medulloblastoma: the current consensus. *Acta Neuropathol.* **123**, 465–472 (2012).
- Cavalli, F. M. G. et al. Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer Cell* **31**, 737–754 (2017).
- Huang, F. W. et al. Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2012).
- Gan, K. A., Carrasco Pro, S., Sewell, J. A. & Fuxman Bass, J. I. Identification of single nucleotide non-coding driver mutations in cancer. *Front. Genet.* **9**, 16 (2018).
- Manser, T. & Gesteland, R. F. Human U1 loci: genes for human U1 RNA have dramatically similar genomic environments. *Cell* **29**, 257–264 (1982).
- Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
- Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-seq data. *Proc. Natl Acad. Sci. USA* **111**, E5593–E5601 (2014).
- Lee, J. H., You, J., Dobrota, E. & Skalniak, D. G. Identification and characterization of a novel human PP1 phosphatase complex. *J. Biol. Chem.* **285**, 24466–24476 (2010).
- Tessemma, M. et al. Differential epigenetic regulation of TOX subfamily high mobility group box genes in lung and breast cancers. *PLoS ONE* **7**, e34850 (2012).
- Kogerman, P. et al. Alternative first exons of *PTCH1* are differentially regulated *in vivo* and may confer different functions to the PTCH1 protein. *Oncogene* **21**, 6007–6016 (2002).
- Sasaki, H., Nishizaki, Y., Hui, C., Nakafuku, M. & Kondoh, H. Regulation of Gli2 and Gli3 activities by an amino-terminal repression domain: implication of Gli2 and Gli3 as primary mediators of Shh signaling. *Development* **126**, 3915–3924 (1999).
- Huard, J. M., Forster, C. C., Carter, M. L., Sicinski, P. & Ross, M. E. Cerebellar histogenesis is disturbed in mice lacking cyclin D2. *Development* **126**, 1927–1935 (1999).
- Kenney, A. M. & Rowitch, D. H. Sonic hedgehog promotes G<sub>1</sub> cyclin expression and sustained cell cycle progression in mammalian neuronal precursors. *Mol. Cell. Biol.* **20**, 9055–9067 (2000).
- Mirzaa, G. et al. *De novo* *CCND2* mutations leading to stabilization of cyclin D2 cause megalencephaly–polymicrogyria–polydactyly–hydrocephalus syndrome. *Nat. Genet.* **46**, 510–515 (2014).
- Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. K. RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* **16**, 413–430 (2016).
- Kim, E. et al. *SRSF2* mutations contribute to myelodysplasia by mutant-specific effects on exon recognition. *Cancer Cell* **27**, 617–630 (2015).
- Mullighan, C. G. et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758–764 (2007).
- Seiler, M. et al. H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. *Nat. Med.* **24**, 497–504 (2018).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

<sup>1</sup>The Arthur and Sonia Labatt Brain Tumour Research Centre, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>2</sup>Developmental and Stem Cell Biology Program, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>3</sup>Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. <sup>4</sup>Informatics and Biocomputing, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. <sup>5</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. <sup>6</sup>Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología, Universidad de Oviedo, Oviedo, Spain. <sup>7</sup>Centro de Investigación Biomédica en Red de Cáncer, Madrid, Spain. <sup>8</sup>CNRS UMR, INSERM, Institut Curie, PSL Research University, Orsay, France. <sup>9</sup>CNRS UMR 3347, INSERM U1021, Université Paris Sud, Université Paris-Saclay, Orsay, France. <sup>10</sup>Tumor Initiation and Maintenance Program, NCI-Designated Cancer Center, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA. <sup>11</sup>Department of Neurosurgery, Division of Pediatric Neurosurgery, Seoul National University Children's Hospital, Seoul, South Korea. <sup>12</sup>Department of Pathology, The Children's Memorial Health Institute, Warsaw, Poland. <sup>13</sup>Centre de Pathologie EST, Groupement Hospitalier EST, Université de Lyon, Bron, France. <sup>14</sup>CNRS UMR5292, INSERM U1028, Centre de Recherche en Neurosciences, Université de Lyon, Lyon, France. <sup>15</sup>Neuro-Oncology Unit, Istituto Giannina Gaslini, Genova, Italy. <sup>16</sup>Division of Pediatric Hematology/Oncology, Mayo Clinic, Rochester, MN, USA. <sup>17</sup>Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA. <sup>18</sup>Department of Pathology, Erasmus University Medical Center, Rotterdam, The Netherlands. <sup>19</sup>Department of Neurology, Erasmus University Medical Center, Rotterdam, The Netherlands. <sup>20</sup>Division of Experimental Medicine, McGill University, Montreal, Quebec, Canada. <sup>21</sup>Department of Anatomical and Cellular Pathology, The Chinese University of Hong Kong, Hong Kong, China. <sup>22</sup>Department of Surgery, The Chinese University of Hong Kong, Hong Kong, China. <sup>23</sup>Department of Pathology, John Hopkins University School of Medicine, Baltimore, MD, USA. <sup>24</sup>Department of Ophthalmology, John Hopkins University School of Medicine, Baltimore, MD, USA. <sup>25</sup>Department of Oncology, John Hopkins University School of Medicine, Baltimore, MD, USA. <sup>26</sup>Department of Neurological Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. <sup>27</sup>Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. <sup>28</sup>Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA. <sup>29</sup>Department of Pediatrics, University of California San Francisco, San Francisco, CA, USA. <sup>30</sup>Department of Neurology, University of California San Francisco, San Francisco, CA, USA. <sup>31</sup>Department of Neurosurgery, Kitasato University School of Medicine, Sagami, Japan. <sup>32</sup>Division of Pediatric Hematology/Oncology, Hospital Pediatría Centro Médico Nacional Century XXI, Mexico City, Mexico. <sup>33</sup>Department of Pathology and Molecular Medicine, Division of Anatomical Pathology, McMaster University, Hamilton, Ontario, Canada. <sup>34</sup>Department of Pathology and Laboratory Medicine, Hamilton General Hospital, Hamilton, Ontario, Canada. <sup>35</sup>Fondazione IRCCS Istituto Nazionale Tumori, Milan, Italy. <sup>36</sup>Winship Cancer Institute, Emory University, Atlanta, GA, USA. <sup>37</sup>Laboratory of Molecular Neuro-Oncology, Department of Neurosurgery, School of Medicine, Emory University, Atlanta, GA, USA. <sup>38</sup>Department of Hematology and Medical Oncology, School of Medicine, Emory University, Atlanta, GA, USA. <sup>39</sup>Department of Neuroscience, Washington University School of Medicine, St. Louis, MO, USA. <sup>40</sup>Department of Pediatrics, Washington University School of Medicine, St. Louis, MO, USA. <sup>41</sup>Department of Pediatrics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. <sup>42</sup>Department of Neurological Surgery, Vanderbilt Medical Center, Nashville, TN, USA. <sup>43</sup>Department of Neurosurgery, Osaka National Hospital, Osaka, Japan. <sup>44</sup>Department of Neurosurgery, Medical and Health Science Centre, University of Debrecen, Debrecen, Hungary. <sup>45</sup>Charbonneau Cancer Institute, University of Calgary, Calgary, Alberta, Canada. <sup>46</sup>Division of Neurosurgery, Centro Hospitalar Lisboa Norte, Hospital de Santa Maria, Lisbon, Portugal. <sup>47</sup>Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal. <sup>48</sup>McGill University and Genome Quebec Innovation Centre, Department of Human Genetics, McGill University, Montreal, Canada. <sup>49</sup>Department of Bioengineering, McGill University, Montreal, Canada. <sup>50</sup>Hopp Children's Cancer Center Heidelberg (KITZ), Heidelberg, Germany. <sup>51</sup>Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ) and German Cancer Consortium (DKTK), Heidelberg, Germany. <sup>52</sup>Department of Pediatric Hematology and Oncology, Heidelberg University Hospital, Heidelberg, Germany. <sup>53</sup>Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. <sup>54</sup>Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>55</sup>Department of Pediatrics, University of California San Diego, San Diego, CA, USA. <sup>56</sup>Department of Surgery, Division of Thoracic and Upper Gastrointestinal Surgery, Faculty of Medicine, McGill University, Montreal, Quebec, Canada. <sup>57</sup>Cancer Research Program, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada. <sup>58</sup>Department of Surgery, Division of Orthopedic Surgery, Faculty of Medicine, McGill University, Montreal, Quebec, Canada. <sup>59</sup>Department of Biochemistry and Molecular Biology, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada. <sup>60</sup>Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. <sup>61</sup>Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada. <sup>62</sup>Division of Haematology/Oncology, Department of Pediatrics, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>63</sup>Department of Cell and Systems Biology, University of Toronto, Toronto, Canada. <sup>64</sup>Division of Neurosurgery, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>65</sup>These authors contributed equally: Hiromichi Suzuki, Sachin A. Kumar. \*e-mail: [mdtaylor@sickkids.ca](mailto:mdtaylor@sickkids.ca)



## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

### Subjects and materials

The study included two large cohorts of medulloblastomas, one from Toronto and the other from the ICGC (Extended Data Fig. 1). The Toronto cohort consisted of 293 cases (whole-genome sequencing, 114 cases; RNA sequencing (RNA-seq) 225 cases; 46 cases overlapped), collected at diagnosis after informed consent was obtained from subjects as part of the Medulloblastoma Advanced Genomics International Consortium. All patient recruitment and tumour-sample collection was approved and in compliance with the ethical regulations of each of the following institutions: The Hospital for Sick Children, Seoul National University Children's Hospital, The Children's Memorial Health Institute, Mayo Clinic, The Chinese University of Hong Kong, John Hopkins University School of Medicine, Seattle Children's Hospital, University of California San Francisco, McMaster University, Erasmus University Medical Center, Kitasato University School of Medicine, Fondazione IRCCS Istituto Nazionale Tumori, Emory University, Osaka National Hospital, Washington University School of Medicine, University of Calgary, Children's Hospital of Pittsburgh, Hospital Pediatria Centro Médico Nacional Century XXI, University of Debrecen, McGill University, Vanderbilt Medical Center, University of Colorado Denver, Istituto Giannina Gaslini and Université de Lyon. The whole-genome sequences consisted of 109 published<sup>3</sup> and 5 unpublished (WNT,  $n = 2$ ; SHH,  $n = 37$ ; group 3,  $n = 26$ ; and group 4,  $n = 49$ ). Samples were obtained as freshly frozen tissue from the time of diagnosis and stored at  $-80^{\circ}\text{C}$  until processed for the purification of nucleic acids. Genomic DNA was isolated by incubation with proteinase K overnight at  $55^{\circ}\text{C}$  followed by 3 sequential phenol extractions and ethanol precipitation. Messenger RNA library construction and sequencing were performed as previously described<sup>24</sup>. The ICGC cohort consisted of 227 cases, which were downloaded from ICGC under accession DACO-1036229.

### Whole-genome sequencing

Whole-genome sequencing was performed at Canada's Michael Smith Genome Science Centre at the BC Cancer Agency, using the Illumina HiSeq 2000/2500 platform as previously described<sup>24</sup>.

### Sequence alignment of whole-genome sequencing data

Whole-genome sequencing reads were aligned to the human reference genome 'hs37d5' by 1000 Genomes Project Phase II, using Burrows-Wheeler aligner (BWA) – MEM version 0.7.8 with the '-T 0' parameter. Duplicates were marked using biobambam v.0.0.148. Sequencing coverages were calculated using GenomonQC software, which was downloaded from Genomon Project, and are shown in Supplementary Table 1.

### Somatic variant calling

Somatic variants were called using eight variant callers: MuTect2<sup>25</sup>, EBCall<sup>26</sup>, Varscan2<sup>27</sup>, Strelka<sup>28</sup>, SomaticSniper<sup>29</sup>, Virmid<sup>30</sup>, Platypus<sup>31</sup> and Seurat<sup>32</sup>.

MuTect2 was run using GATK v.3.5.0 with the default setting. Candidate variants were discarded if the variants were detected in a panel of normal that was made by MuTect2 with '-artefact\_detection\_mode' and GATK 'CombineVariants' function with '-minN 2'. EBCall v.0.2.1 was run with the default setting. We used the following criteria: requiring  $P$  value (by EBCall)  $< 10^{-3}$ , variant reads in tumour  $\geq 2$  and variant reads in normal  $\leq 1$ . Varscan2 v.2.4.3 was run with parameters '-strand-filter 1 -min-var-freq 0.08'. The results were filtered using the 'fpfilter' function with the option '-dream3-settings'. Strelka v.1.0.15 was run with default parameters. Virmid v.1.1.0 was run with the option '-q 10'. Somatic Sniper v.1.0.5.0 was run with the parameters '-Q 15 -q 1 -G -L'

and the results were filtered by the author's recommended filter using bam-readcount. The candidates with more than 0.03 of variant allele frequency in the matched normal sample were discarded. Platypus v.0.8.1 was run with a default setting. Detected variants that passed the standard Platypus filtering criteria or showed 'allele bias' were used. We used the following additional criteria: requiring likelihood(reference allele)/likelihood(variant allele)  $< 10^{-5}$  in tumour, likelihood(variant allele)/likelihood(reference allele)  $< 10^{-5}$  in matched control, variant reads in tumour  $\geq 2$  and variant reads in normal  $\leq 1$ . Seurat v.2.5 was run with the option '-indels'. We used variants that were called by at least two callers. Obtained results were filtered by  $\leq 2$  variant reads in the matched normal control, calculated using the realignment function of Genomon-MutationFilter v.0.2.1. Variants were annotated using ANNOVAR<sup>33</sup>. The correlations of U1 and U11 snRNA mutations with other somatic events were analysed using R package 'Epi' v.2.30. Asymptotic  $P$  values from odds-ratio tests were calculated using the twoby2 function, followed by Benjamini-Hochberg adjustment for multiple testing.

### Copy-number calling for whole-genome sequencing

Copy-number alterations were detected using Control-FREEC v.10.3 with the following parameters: breakPointType = 4, ploidy = "2,3,4", step = 10000, window = 50000 (ref.<sup>34</sup>).

### Variant calling of U1 and U11 snRNA genes

To explore mutations in low-mappability regions, we first picked up reads from whole-genome sequencing data on U1 and U11 snRNA genes and pseudogenes using samtools and biobambam. To accept multimapping, we used STAR aligner<sup>35</sup>. To prevent gaps, we set the setting with '-scoreGap -20 -alignEndsType EndToEnd'. Mutations were called by EBCall with the same setting as for whole-genome sequencing, except for the acceptance of secondary alignment. We used the following criteria: requiring  $P$  value (by EBCall)  $< 10^{-3}$ , variant reads in tumour  $\geq 4$  and variant reads in matched control  $\leq 1$ .

To evaluate the exact loci of variant reads and multiple mutations of U1 snRNAs, we mapped variant reads to case-specific references. First, we extracted all variant reads of U1 snRNA mutations (r.3A>G) with mate-paired reads. Then, we constructed a case-specific reference that included U1 snRNA hotspot mutations (r.3A>G) and case-specific germline variants detected from extracted variant reads, using the samtools mpileup function. Variant reads were mapped again on the case-specific reference, using bwa-mem with the same setting as was used in the whole-genome sequencing analysis. Using .bam files with the case-specific reference, we called variants in flanking regions of the U1 snRNA hotspot mutation (r.3A>G), using samtools mpileup function to evaluate multiple mutations. No samples had recurrent variant reads. Therefore, we concluded that U1 snRNA mutation occurs in one allele. To interpret the mutated genes, we extracted consecutive consensus sequences of upstream U1 snRNA sequences with two or more supported reads. Then, the consensus sequence was mapped using BLAST software to U1 snRNA genes and pseudogenes with 1,000-bp upstream sequences from the hg19 reference. Because there were many variants and a high level of similarity in the upstream sequences, we cannot detect the exact positions of mutated reads except for *RNVUI-18* mutations. Therefore, we classified U1 snRNA mutations into (1) RNU1 genes (*RNU1-1*, *RNU1-2*, *RNU1-3* or *RNU1-4*), (2) *RNVUI-18* and (3) RNU1 pseudogenes (*RNU1-27P* or *RNU1-28P*) on the basis of the similarity of sequences of the flanking region. Finally, we performed a manual review of detected mutations with the Integrative Genome Viewer<sup>36</sup>. Detected mutations are shown in Supplementary Tables 2–4.

### Secondary structure of U1 and U11 snRNAs

The conservation scores of U1 (RF00003) and U11 (RF00548) snRNAs were downloaded from Rfam<sup>37</sup>. U1 and U11 sequences of other species were downloaded from seed sequences from Rfam. The secondary structures are described on the basis of the consensus structure in Rfam using

VARNA software<sup>38</sup>. U2-type intron and U12-type intron sequences were downloaded from SpliceRack<sup>39</sup>.

### rhAmp genotyping

Genomic DNA from primary tumours was tested using custom rhAmp single-nucleotide polymorphism (SNP) assays (Integrated DNA Technology). In brief, locus- and allele-specific primers were generated individually for RNU1\_batch (*RNU1-1*, *RNU1-2*, *RNU1-3*, *RNU1-4* and *RNU1-18*) and RNU1\_pseudo (*RNU1-27P* and *RNU1-28P*). Assays were run in technical triplicate in 5 µl volume (DNA concentration was at least 5 ng/µl), with control gBlocks for wild-type, mutant and heterozygous genotypes. The reporter mix used Yakima Yellow (mutant) and FAM (wild-type) dyes, as well as ROX dye as a passive reference. Plates were read on the StepOnePlus (Applied Biosystems) RT-PCR machine, and genotypes were called using the StepOne v.2.3 software. The primer sequences are available in Supplementary Table 19.

### RNA-seq

Sequencing reads were mapped by STAR v.2.5.1b on fasta, which includes the human reference genome hs37d5 by 1000 Genomes Project Phase II, spike-in sequences of profile C1\_2 ERCC spike-in concentrations used for C1 fluidigm and Caltech profile 3 spike-ins by Encyclopedia of DNA elements (ENCODE), with the options ‘-outFilterMultimapNmax 20 -alignSJoverhangMin 8 -alignMatesGapMax 200000 -alignIntronMax 200000 -alignSJDBoverhangMin 10 -alignSJstitchMismatchNmax 5-1 55 -outSAMmultNmax 20 -twopassMode Basic’<sup>35</sup>. Mapping results are shown in Supplementary Table 20.

### Intron-centric alternative splicing analysis

Intron-centric alternative splicing analysis was performed using LeafCutter<sup>11</sup>. LeafCutter is an annotation-free quantification method. Intron clustering was run with minimum required reads = 50 and max\_intron = 500,000. LeafCutter was run with the option ‘-g 0’. Each of the 30 cases of each subtype of SHH medulloblastoma was compared with samples of the other subtypes, 5 adult brains and 4 fetal brains, using the default settings. Cases of SHHα with U1 snRNA r.3A>G mutation (*n* = 13) were compared with cases of SHHα with wild-type U1 snRNA (*n* = 39). The results we obtained were filtered by the *q* value of each cluster < 0.01, in cases in which at least one absolute effect size calculated by LeafCutter was more than 1.5. Each event was annotated by LeafViz with GENCODE v.19 .gtf file. Events with unknown strand directions were not analysed. Logo sequences were built using R package ‘ggseqlogo’ v.0.1<sup>40</sup>. Statistical analyses comparing sequences were performed by  $\chi^2$  test. The adjusted standardized residual was calculated by Haberman’s method. We selected cryptic 5’ splicing events with a C base at the sixth base in the intron. Subsequently, we further prioritized alternatively spliced genes that have previously been reported as recurrent genetic aberrations in SHH medulloblastoma<sup>3,41</sup>, are transcriptionally upregulated or downregulated in both the SHHα and SHHδ subtypes<sup>7</sup>, or registered as tier 1 in Cancer Gene Census.

*t*-distributed stochastic neighbour embedding (*t*-SNE) analysis was performed using the R package ‘Rtsne’ v.0.13. Analysed events were chosen with the following: (1) significant events in at least one SHH subtype; and (2) lengths of the cluster of junction reads are same among all subtypes. Per cent spliced in was calculated by the number of junction reads of alternative splicing events, divided by the total number of junction reads in a cluster. *t*-SNE was run with a default setting along with 3 WNT, 20 group 3 and 22 group 4 medulloblastomas, which were used in a previous study<sup>42</sup>.

### Exon-centric alternative splicing analysis

Exon-centric alternative splicing analysis was performed using rMATS v.4.0.1<sup>12</sup>. rMATS was run with default setting with GENCODE v.19 for alternative 3’ splice site, alternative 5’ splice site, retained intron and skipped exon. We filtered the events with FDR < 0.01 and change of

splicing inclusion (calculated by rMATS) > 0.05. The Sashimi plots were described using MISO v.0.5.4<sup>43</sup>.

### Gene-set enrichment analysis of nonsense-mediated decay

We counted reads using GENCODE v.19 .gtf file and htseq v.0.6.0 with the setting ‘-stranded reverse -m union’. Differential expression analysis was performed using DESeq2 v.1.16.1 with the default setting, after extracting genes expressed at >5 counts per million in at least 20% of cases. We performed two comparisons: SHHδ with mutant U1 snRNA (*n* = 30) versus other SHH subtypes with wild-type U1 snRNA (*n* = 90) and SHHα with mutant U1 snRNA (*n* = 13) versus SHHα with wild-type U1 snRNA (*n* = 39). Gene-set enrichment analysis (GSEA) for differentially expressed genes was performed using pre-ranked gene lists ordered by  $-\log_{10}$  (*P*value) multiplied by +1 for upregulation or -1 for downregulation with gsea v.3.0. We used two datasets for a pathway of nonsense-mediated decay, ‘GO nuclear transcribed mrna catabolic process nonsense mediated decay’ from the C5 gene set and ‘Reactome nonsense mediated decay enhanced by the exon junction complex’ from the C2 gene set.

### TP53 mutation status

Germline mutations of *TP53* were analysed using EBCall v.0.2.1. EBCall was run with the default setting. We used the following criteria: requiring *P*value (by EBCall) <  $10^{-3}$ , 90% posterior quantile calculated by EBCall > 0.3. The results were annotated using ANNOVAR.

The mutation call from RNA-seq was run using GATK v.3.8.0. Adding read groups and flagging duplicate reads were performed using Picard tool v.2.18.0. Then, we split reads into exon segments using GATK with the setting ‘-rf ReassignOneMappingQuality -RMQF 255 -RMQT 60 -U ALLOW\_N\_CIGAR\_READS’. Base recalibration was performed using GATK. Mutation call was performed using the ‘HaplotypeCaller’ function of GATK with the setting ‘-dontUseSoftClippedBases -stand\_call\_conf 20.0’. Variants were filtered using the ‘VariantFiltration’ function of GATK with the setting ‘-window 35 -cluster 3 -filterName FS -filter “FS > 30.0” -filterName QD -filter “QD < 2.0”’. The variants were discarded if they were also detected in any RNA-seq data generated from nine normal brain samples (five adult brains and four fetal brains). Sanger sequencing was as performed in a previous study<sup>44</sup>. We discarded the mutations that showed a frequency of 0.01 or higher in 1000 Genomes v5b or ESP-6500, or dbSNP138.

### Survival analysis

Overall survival and progression-free survival were evaluated using the log-rank with R package ‘survival’ v.2.40.1. Overall survival was defined as the time from date of surgery to death or date of last follow-up, and progression-free survival as the time from date of surgery to first event (progression or relapse) or date of last follow-up.

### Pan-cancer analysis

We analysed 2,442 cases of cancer across 36 tumour types from ICGC. The hotspot mutations were analysed with the method described in ‘Variant calling of U1 and U11 snRNA genes’, except for the use of the mapping tool. For pan-cancer data, we used bowtie aligner instead of STAR<sup>45</sup>.

### SNP6 copy-number analysis

Array files were downloaded from the Gene Expression Omnibus (GEO) under accession number GSE37385, and the relevant Affymetrix SNP6 arrays were extracted. Affymetrix Power Tools v.1.18.2 was used to process and normalize the probe intensities to generate log *R* ratio (LRR) and B-allele frequency (BAF) using the PennCNV-Affy pipeline<sup>46</sup>. The affywg6.hg19.pfb file was used to map the probes onto the hg19 genome. All other parameters were left as default.

The resulting probe-level LRR and BAF values were taken into ASCAT v.2.4.3<sup>47</sup>. GC wave correction was then performed, followed by predicting germline genotypes; this finally led to running the ASCAT algorithm to determine the copy-number values for each genomic region, as well as

# Article

the overall ploidy and purity of the sample. Samples with a model fit that was less than 80% failed their ASCAT processing stage. log ratios for each segment were calculated by using the copy number of each segment, as well as the average ploidy of the sample, according to the equation:

$$\text{ratio} = \log_2 \left( \frac{\text{copy number}}{\text{ploidy}} \right)$$

Adjacent segments with log ratios that differed by less than 0.25 were then merged using their size-weighted mean:

$$\text{new ratio} = \frac{\text{length1} \times \text{ratio1} + \text{length2} \times \text{ratio2}}{\text{length1} + \text{length2}}$$

Copy-number states were assigned to each segment on the basis of their log ratio and their ploidy values, according to the Supplementary Table 21. Broad copy-number changes are defined as occurring in 75% or more of chromosome arm in size. Focal copy-number variants were analysed using GISTIC v.2.0.23<sup>48</sup>. GISTIC was run with the settings ‘-ta 0.25 -td 0.3 -js 10 -brlen 0.7 -gcm “extreme” -armpeel’.

## PCR with reverse transcription and qPCR analysis

RNA was obtained for samples from 18 patients that had fragments per kilobase of transcript per million mapped reads (FPKM) values of more than 2 for targeted genes from our larger cohort (6 SHHα with wild-type U1 snRNA, 6 SHHα with mutant U1 snRNA and 6 SHHδ with mutant U1 snRNA). cDNA was synthesized using SuperScript III (Thermo Fisher 18080400). PCRs were performed with cDNA and Taq polymerase using 35 cycles, and products run on a 2% agarose gel. qPCRs were performed using SYBR-Green with ROX (Thermo Fisher 11744500), two steps at 35 cycles. Calculation of  $\Delta\Delta C_T$  was done comparing the expression of the mutant isoform to the wild-type isoform. The primer sequences are available in Supplementary Table 19.

## Generation of a lentiviral vector for the expression of U1 snRNA r.3A>G

The pLKO.1-puro U6 sgRNA BfuAI stuffer lentiviral vector (Addgene no. 50920) was modified by removing the internal U6 promoter (between NdeI and EcoRI), and it was replaced by the U1 snRNA locus, including 393 bases of internal native U1 promoter, the U1 sequence, and 39 bases of 3′-flanking region using the following oligonucleotides (5′-GTCC AGAATTCTTGGCGTACAGTCTGTTTTTG and 5′-CTATCATATGTAAGGACCAGCTTCTTTGGGA). The PCR products were digested with NdeI and EcoRI, and cloned in the modified pLKO.1 plasmid. The r.3A>G mutation was introduced by site-directed mutagenesis. All plasmids were verified by Sanger sequencing.

## Exogenous expression of U1 snRNA r.3A>G mutation

Cell lines used were human embryonic kidney 293T (HEK-293T) from American Type Culture Collection (ATCC) (CRL-3216). Cell stocks were mycoplasma-tested before U1 vector transfection. HEK-293T cells were grown in DMEM, 10% FBS and 1% PSG. For exogenous expression of U1 snRNA, HEK-293T cells ( $5 \times 10^6$  cells) were cultured in 10-cm plates and transfected using Lipofectamine Plus (Invitrogen) with 2 μg of either pLKO.1-U1wt (containing the wild-type U1 snRNA locus) or pLKO.1-U1r.3A>G (containing the r.3A>G mutation) in duplicate. Twelve hours after transfection, the medium was replaced with complete medium, and 48 h later the total RNA was extracted with the Trizol method.

## Verification of the expression of U1 snRNA r.3A>G mutation

Rapid amplification of cDNA ends (RACE) was performed using 1 μg of total RNA from HEK-293T cells transfected with either pLKO.1-U1wt or pLKO.1-U1r.3A>G following the recommendations of the manufacturer (Sigma-Aldrich 3353621001), and the following specific oligonucleotides (U1-RACE-SP1: 5′-CAGGGGAAAGCGCGAACGCACT and U1-RACE-SP2:

5′-CCCCTACCACAAATTATGC). A single amplification band of the expected size (160 bp) was excised from the gel, purified and sequenced with the internal oligonucleotide U1-RACE-SP2.

## Sequence analyses of exogenous expression analysis

Messenger RNA library construction was performed based on oligo dT-based mRNA isolation using NEBNext Poly(A) mRNA Magnetic Isolation Module. RNA sequencing was performed on NextSeq 550 using 100-bp paired-end mode. Mapping and intron clustering were performed with the methods described in ‘RNA-seq’ and ‘Intron-centric alternative splicing analysis’. LeafCutter was run with the option ‘-g 0 -i 2’ and the obtained results were filtered by a *q* value of each cluster < 0.1, in which at least 1 absolute effect size calculated by LeafCutter was more than 1.5.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

Sequencing data have been deposited in the European Genome-Phenome Archive (EGA) and Gene Expression Omnibus (GEO): RNA-seq (EGAD00001001899 and EGAD00001004958), whole-genome sequences (EGAD00001003125 and EGAD00001004347) and RNA-seq of exogenous expression analyses (GSE128005). Materials used in this study are available from the following: GENCODE ([ftp://ftp.ebi.ac.uk/pub/databases/genencode/Gencode\\_human/release\\_19/gencode.v19.anotation.gtf.gz](ftp://ftp.ebi.ac.uk/pub/databases/genencode/Gencode_human/release_19/gencode.v19.anotation.gtf.gz)), ICGC (<https://icgc.org/>), hs37d5 reference ([ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2\\_reference\\_assembly\\_sequence](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence)), Burrows–Wheeler aligner (bwa) (<http://bio-bwa.sourceforge.net/>), Mutect2 (<https://software.broadinstitute.org/gatk/>), EBCall (<https://github.com/friend1ws/EBCall>), VarScan2 (<http://dkoboldt.github.io/varscan/>), Strelka (<https://github.com/Illumina/strelka>), SomaticSniper (<http://gmt.genome.wustl.edu/packages/somatic-sniper/>), Virmid (<https://sourceforge.net/p/virmid/wiki/Home/>), Platypus (<http://www.well.ox.ac.uk/platypus>), Seurat (<https://sites.google.com/site/seuratsomatic/home>), ENCODE (<https://www.encodeproject.org/>), PennCNV (<http://penncnv.openbioinformatics.org/en/latest/>), Database of Genomic Variants (<http://dgv.tcag.ca/dgv/app/home>), Genomon Project (<https://github.com/Genomon-Project>), SpliceRack (<http://katahdin.mssm.edu/splice/index.cgi?database=spliceNew>) and GEO (<https://www.ncbi.nlm.nih.gov/geo/>)

24. Morrissey, A. S. et al. Divergent clonal selection dominates medulloblastoma at recurrence. *Nature* **529**, 351–357 (2016).
25. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
26. Shiraishi, Y. et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41**, e89 (2013).
27. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
28. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
29. Larson, D. E. et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
30. Kim, S. et al. Virmid: accurate detection of somatic mutations with sample impurity inference. *Genome Biol.* **14**, R90 (2013).
31. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
32. Christoforides, A. et al. Identification of somatic mutations in cancer through Bayesian-based analysis of sequenced genome pairs. *BMC Genomics* **14**, 302 (2013).
33. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
34. Boeva, V. et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2012).
35. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
36. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
37. Kalvari, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **46**, D335–D342 (2018).
38. Darty, K., Denise, A. & Ponty, Y. VARNAs: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975 (2009).

39. Sheth, N. et al. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* **34**, 3955–3967 (2006).
40. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
41. Northcott, P. A. et al. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature* **488**, 49–56 (2012).
42. Pei, Y. et al. HDAC and PI3K antagonists cooperate to inhibit growth of MYC-driven medulloblastoma. *Cancer Cell* **29**, 311–323 (2016).
43. Katz, Y., Wang, E. T., Airolidi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
44. Zhukova, N. et al. Subgroup-specific prognostic implications of TP53 mutation in medulloblastoma. *J. Clin. Oncol.* **31**, 2927–2935 (2013).
45. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
46. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
47. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
48. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).

**Acknowledgements** M.D.T. is supported by the NIH (R01CA148699 and R01CA159859), The Pediatric Brain Tumour Foundation, The Terry Fox Research Institute, The Canadian Institutes of Health Research, The Cure Search Foundation, b.r.a.i.n.child, Meagan's Walk, Genome Canada, Genome BC, Genome Quebec, the Ontario Research Fund, Worldwide Cancer Research, V-Foundation for Cancer Research and the Ontario Institute for Cancer Research, through funding provided by the Government of Ontario. M.D.T. is also supported by a Canadian Cancer Society Research Institute Impact grant and by a Stand Up To Cancer (SU2C) St. Baldrick's Pediatric Dream Team Translational Research Grant (SU2C-AACR-DT1113) and SU2C Canada Cancer Stem Cell Dream Team Research Funding (SU2C-AACR-DT-19-15) provided by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, with supplementary support from the Ontario Institute for Cancer Research through funding provided by the Government of Ontario. SU2C is a programme of the Entertainment Industry Foundation administered by the American Association for Cancer Research. M.D.T. is also supported by the Garron Family Chair in Childhood Cancer Research at the Hospital for Sick Children and the University of Toronto. E.G.V.M. is supported by the NIH (R01-NS096236 and R01CA235162) and the CURE Childhood Cancer Foundation. X.S.P. is supported by Ministerio de Economía y Competitividad (MINECO) (SAF2013-45836-R). A.K. was supported by 2017-1.2.1-NKP-2017-00002 National Brain Research Program NAP 2.0. M.L.G.

is supported by AIRC (Italian Association for Cancer Research) and by Fondazione Berlucchi. H.S. is a recipient of a Research Fellowship (Astellas Foundation for Research on Metabolic Disorders). S.A.K. is a recipient of funding from the Restrcomp Research Fellowship (SickKids Research Institute) and the MD/PhD Studentship Award (Canadian Institute of Health Research). A.D.-N. is a recipient of the Department of Education of the Basque Government (PRE\_2017\_1\_0100). J.R. is supported by Genome Canada Genome Technology Platform Grant 12505, Canada Foundation for Innovation Project 33408. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics and on the Niagara supercomputer at the SciNet HPC Consortium. SciNet is funded by the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto. The authors would also like to thank M. Kamal and J. Loukides, and recognize the Labatt Brain Tumour Research Centre.

**Author contributions** M.D.T. led the study. H.S., S.S. and S.D.B. performed whole-genome sequencing analysis (Figs. 1, 2, Extended Data Figs. 1–3). F.M.G.C., N.G., J.R. and A.S.M. contributed to the pre-processing of RNA-seq data. H.S. and H.F. contributed to SNP6 copy-number analyses (Extended Data Figs. 1, 4c, d). H.S., S.S., F.M.G.C., I.S. and J.Z. contributed to RNA expression analyses (Extended Data Fig. 7a). H.S., S.S., I.S., A.F., S.D.B. and O.A. contributed to alternative splicing analyses (Fig. 4a–d, Extended Data Figs. 5–10). H.S. and V.R. performed clinical analysis (Fig. 3, Extended Data Fig. 4e–i). A.G. and M.A.M. helped with bioinformatics analyses and provided expert advice. S.A.K., P.D.A., K.J. and M.C.V. performed real-time PCR and qPCR analyses (Fig. 4e, f, Extended Data Figs. 9d, 10e, i). S.A.K., A.D.-N., A.G.-F., P.D.A., K.J., I.S., N.A., D.P., A.M., J.W., W.D., R.J.W.-R. and X.S.P. contributed to exogenous expression experiments (Extended Data Fig. 7b–d). S.A.K., K.J. and I.S. performed rhAMP SNP experiments (Figs. 1, 3a–c, Extended Data Figs. 1, 4a, b). P.S. and B. Luu contributed to the collection and processing of human tissue samples. C.D., X.W., R.J.W.-R., L.G., X.H., X.S.P., J. A. Chan and L.S. provided expert advice for experiments. S.-K.K., W.A.G., A.J., M.F.-M., M.L.G., A.A.N.-R., C.G., J.M.K., P.J.F., N.J., H.-K.N., W.S.P., C.G.E., I.F.P., J.M.O., W.A.W., T.K., E.L.-A., B. Lach, M.M., E.G.V.M., J.B.R., R.V., L.B.C., N.K., A.K., L.B., J. A. Calarco, C.C.F., S.M.P., L.G. and D.M. provided patient material and helped design the study. H.S., S.A.K., S.S., J. A. Calarco, L.S. and M.D.T. prepared the manuscript and figures.

**Competing interests** The authors declare no competing interests.

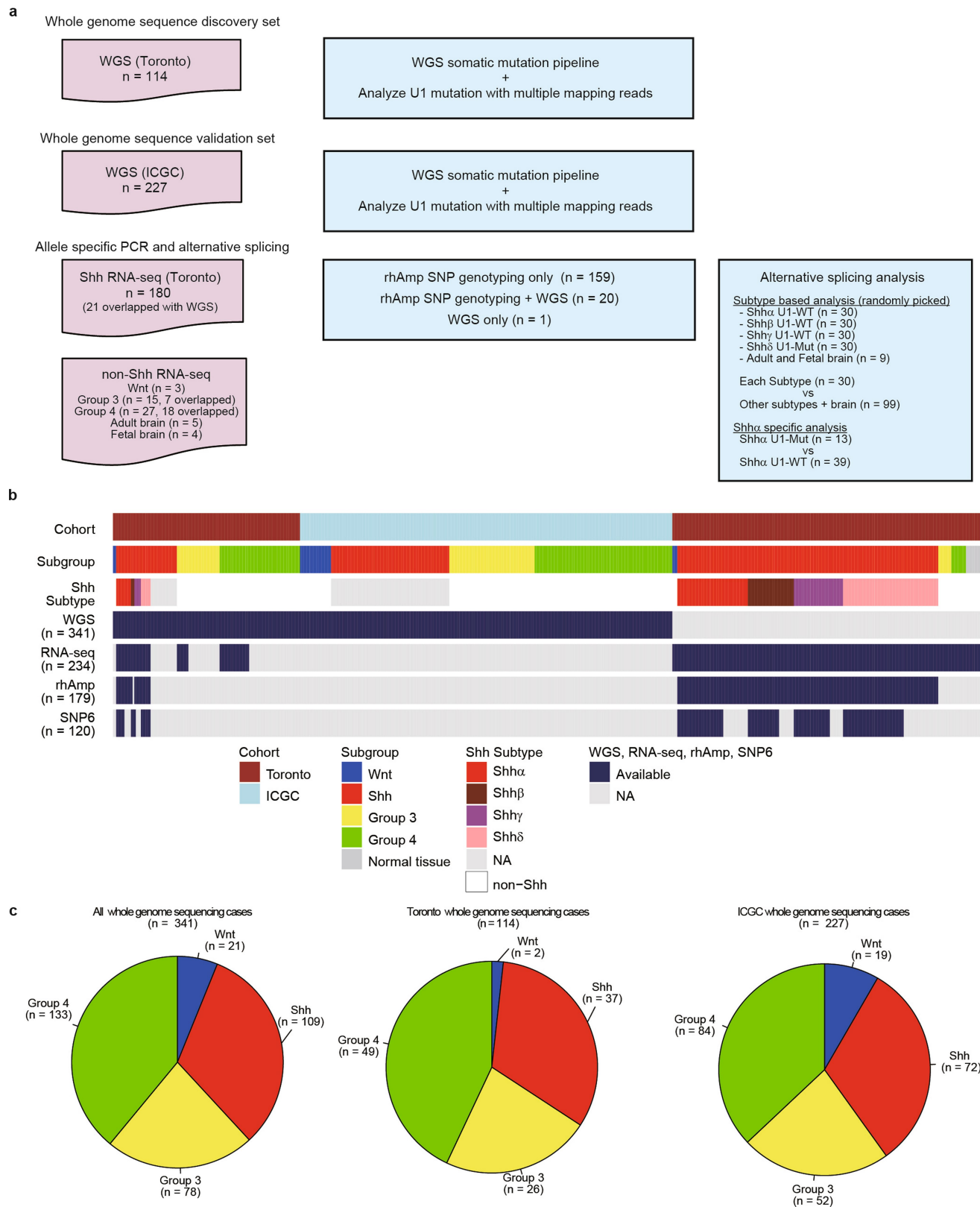
#### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1650-0>.

**Correspondence and requests for materials** should be addressed to M.D.T.

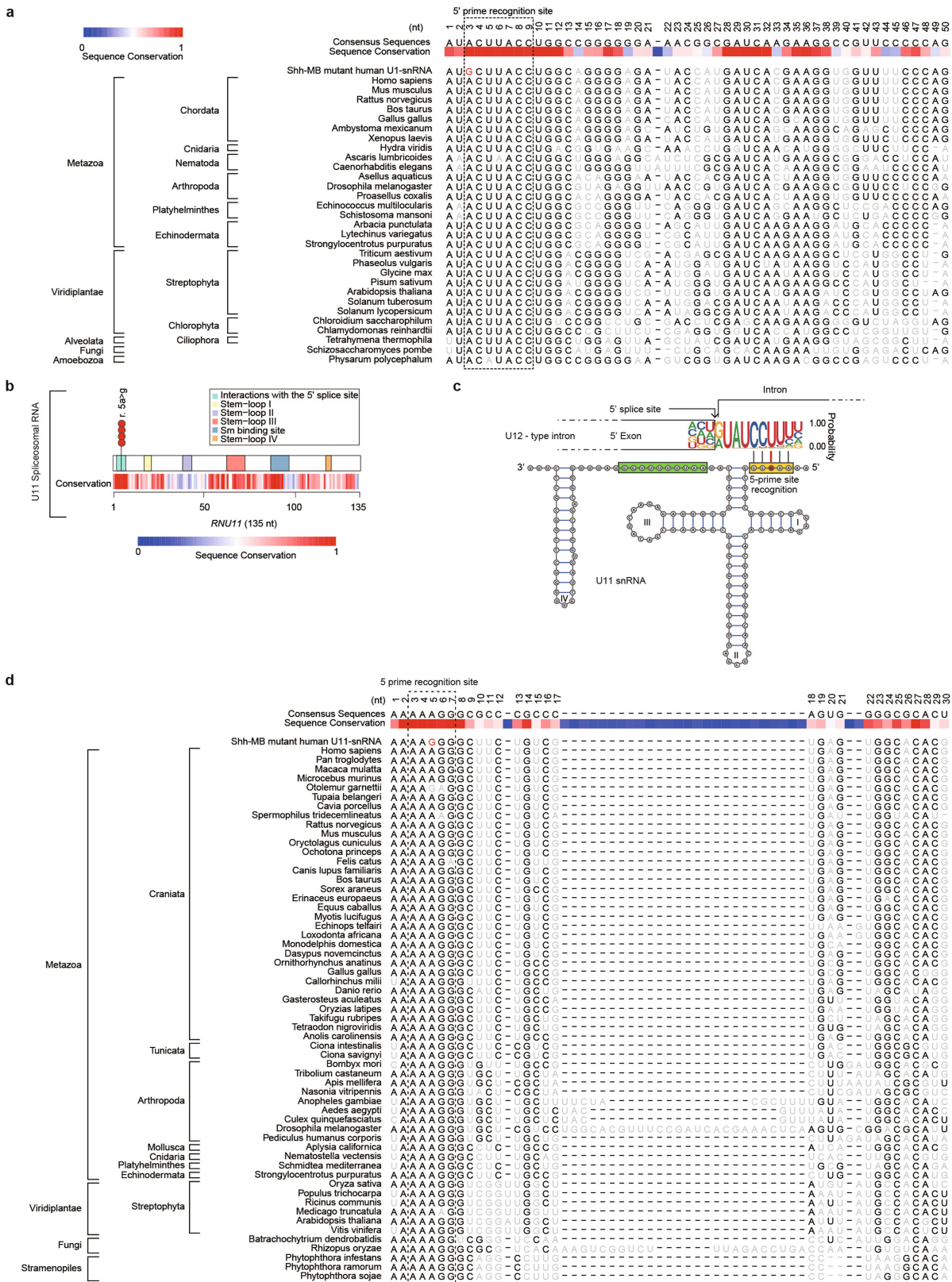
**Peer review information** *Nature* thanks Rotem Karni, Brandon Wainwright and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Overview of analysed cohorts and methods. a**, The detection methods for U1 snRNA mutations by each cohort, and comparison methods for alternative splicing analysis. **b**, Cohort specification. **c**, Subgroup distribution of whole-genome sequencing (WGS) cohorts.

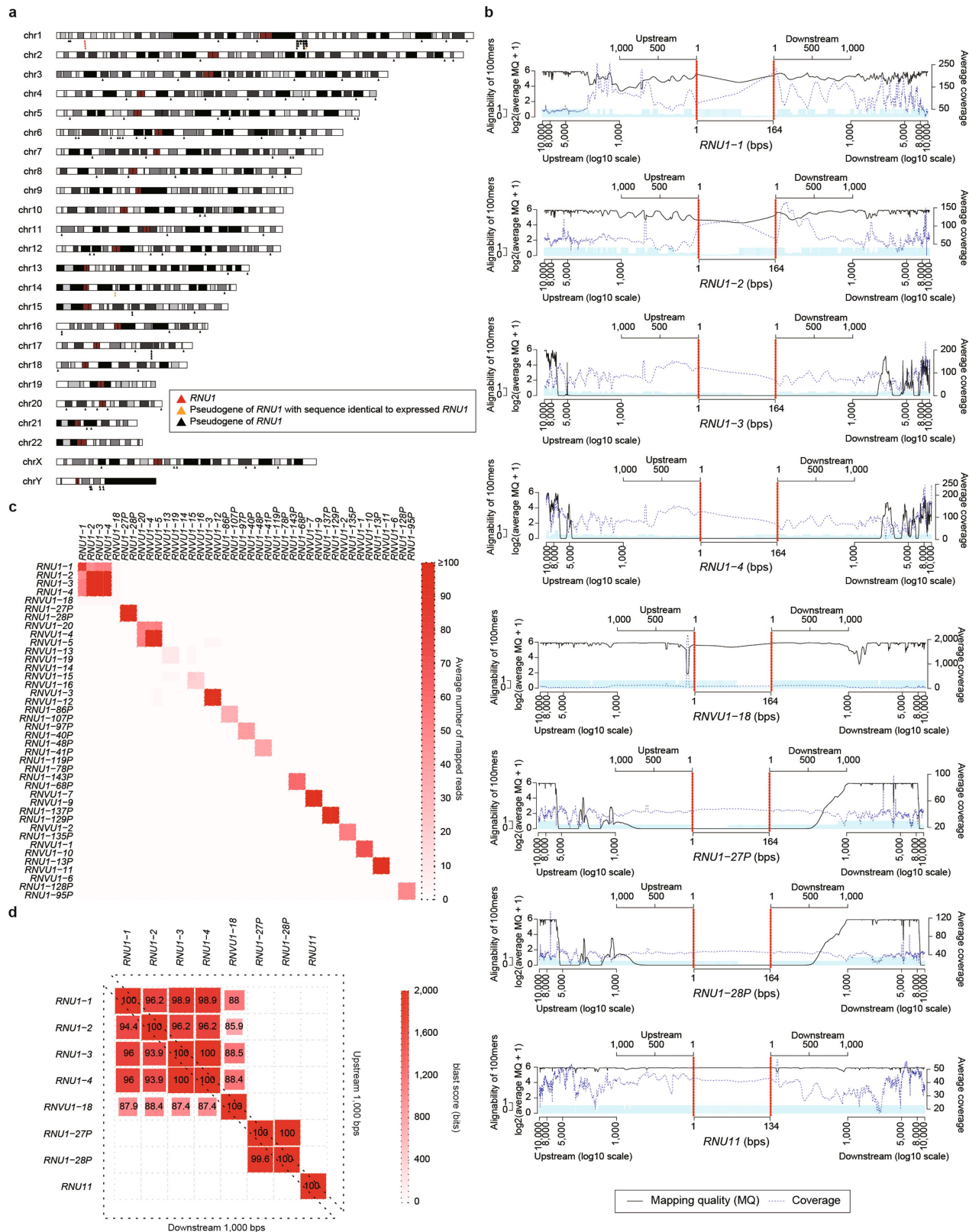




Extended Data Fig. 2 | See next page for caption.

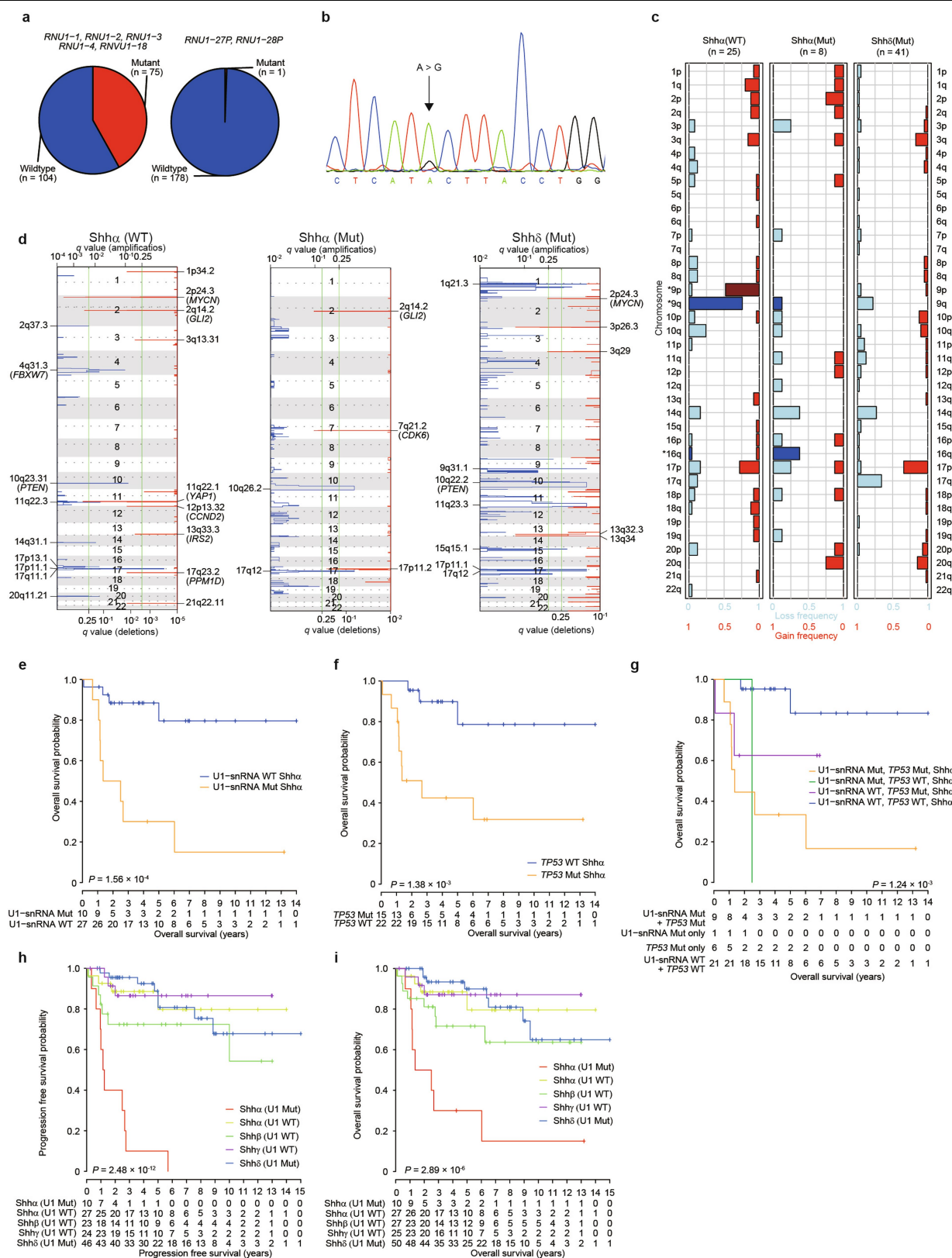
**Extended Data Fig. 2 | U11 snRNA mutations, and conservation of U1 and U11 snRNA genes across evolution.** **a**, Seed sequences of the U1 snRNA obtained from the Rfam database demonstrate high-level conservation across a variety of eukaryotic species, particularly at the site of the SHH medulloblastoma mutation. The consensus sequence and first 50 nucleotides of reference sequences are included for comparison. Grey indicates nucleotide differences, and red identifies the SHH medulloblastoma hotspot mutation. **b**, Cartoon illustrating the number of somatic mutations in the U11 snRNA genes. Sequence conservation scores for U11 snRNA, as determined using the Rfam database. **c**,

Secondary structure of the mutant U11 snRNA. The red circle identifies the location of the hotspot mutation. The yellow and green rectangles indicate the 5' splice-site recognition site and the Sm protein-binding site, respectively. Numerals I to IV indicate stem loops. **d**, Seed sequences of the U11 snRNA obtained from the Rfam database demonstrate high-level conservation across a variety of eukaryotic species, particularly at the site of the SHH medulloblastoma mutation. The consensus sequence and first 30 nucleotides of reference sequences are included for comparison. Grey indicates nucleotide differences, and red identifies the SHH medulloblastoma hotspot mutation.



**Extended Data Fig. 3 | High levels of genomic conservation surrounding human U1 snRNAs complicate the specific PCR amplification of any individual locus.** **a**, Genomic locations of the 4 expressed U1 snRNA genes (on chromosome 1p, red) and 136 pseudogenes across the *Homo sapiens* genome, as indicated. Three pseudogenes with sequences that are identical (hg19) to the expressed U1 snRNA genes are indicated in orange. **b**, Average mapping quality of bwa-mem and coverage of each of the expressed U1 and U11 snRNA genes from whole-genome sequencing of germline samples from patients with medulloblastoma are illustrated ( $n = 341$  patients). Blue bars represent the alignability of 100-mers using the GENome Multitool (GEM) mapper from

ENCODE and Centre for Genomic Regulation (CRG). Regions that are >1,000 bases upstream and downstream are shown on a  $\log_{10}$  scale. Red bar indicates the gene body. **c**, Average number of multimapped reads overlapped for each gene pair using STAR aligner. The heat map shows the average number of mapped reads across the whole-genome sequencing of germline samples from patients with medulloblastoma ( $n = 341$  patients). **d**, Sequence similarity of U1 snRNA genes, U1 snRNA pseudogenes with 164 identical base pairs and the U11 snRNA gene. The numbers in each square and heat map indicate identity scores and bit scores calculated using blast software. A blank square indicates that no hit was found.

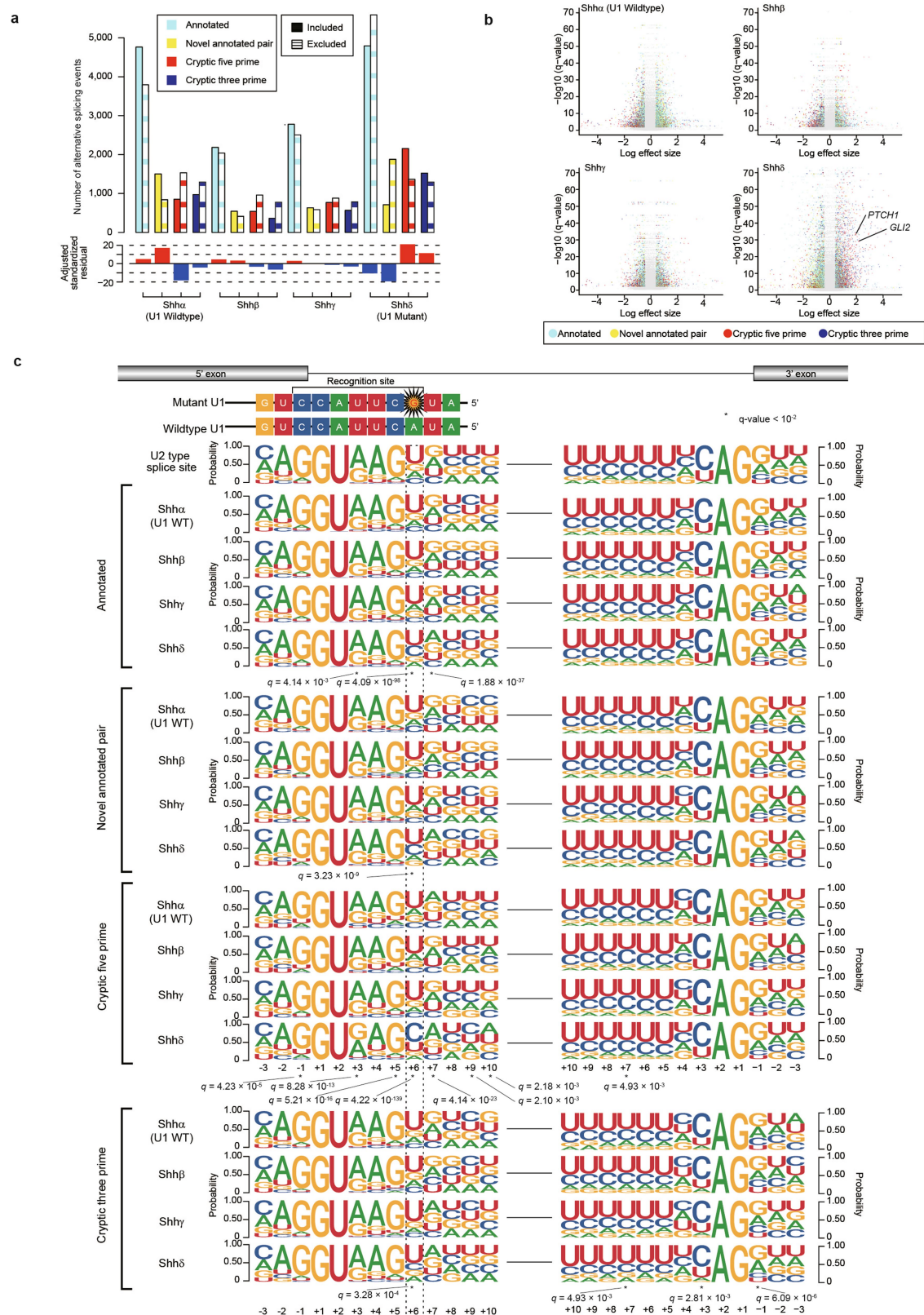


**Extended Data Fig. 4** | See next page for caption.



**Extended Data Fig. 4 | Allele-specific rhAmp SNP PCR of RNU1 loci, copy-number changes in SHH medulloblastoma with mutant U1 snRNA versus wild-type U1 snRNA, and prognostic analysis.** **a**, The frequency of any U1 snRNA mutation in the RNU1\_batch primer set (*RNU1-1*, *RNU1-2*, *RNU1-3*, *RNU1-4* and *RNVU1-18*) (left) and the RNU1\_pseudo primer set (*RNU1-27P* and *RNU1-28P*) (right). **b**, Hotspot mutations of *RNU1-27P* or *RNU1-28P* U1 snRNA pseudogenes, as confirmed by Sanger sequencing. **c**, Broad copy-number aberrations in SHHα with wild-type U1 snRNA ( $n = 25$ ), SHHα with mutant U1 snRNA ( $n = 8$ ) and SHHδ with mutant U1 snRNA ( $n = 41$ ). Dark blue and dark red bars, as well as asterisks, identify statistically significant regions, comparing SHHα with mutant versus wild-type U1 snRNA ( $P < 0.05$ , two-sided Fisher's exact test). **d**, Significant focal copy-number aberrations in SHHα with wild-type U1 snRNA ( $n = 25$ ), SHHα with mutant U1 snRNA ( $n = 8$ ) and SHHδ with mutant U1 snRNA ( $n = 41$ ) illustrate

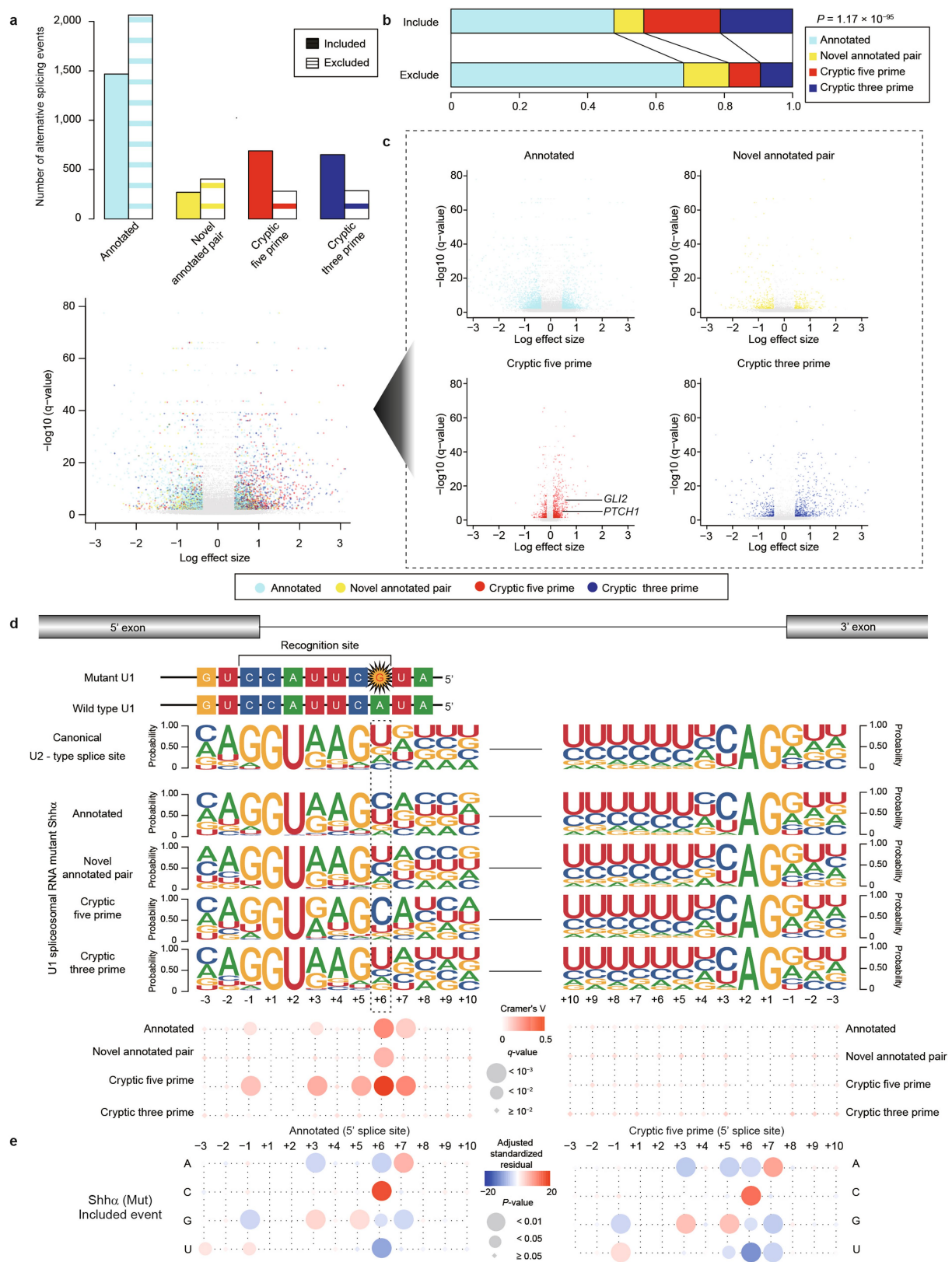
significant genomic differences between cases with wild-type and mutant U1 snRNA. Candidate target genes within the corresponding loci are indicated.  $q$  values were calculated using GISTIC (Methods). **e–g**, Overall survival of patients with SHHα, stratified by mutational status of U1 snRNA mutation ( $n = 10$  for mutant,  $n = 27$  for wild type) (**e**), *TP53* ( $n = 15$  for mutant,  $n = 22$  for wild type) (**f**) or both ( $n = 9$  for both mutant,  $n = 1$  for U1 snRNA mutation only,  $n = 6$  for *TP53* mutation only,  $n = 21$  for both wild type) (**g**).  $P$  values were determined using the two-sided log-rank test. **h, i**, Progression-free survival (**h**) and overall survival (**i**) stratified by U1 snRNA mutation and SHH subtype ( $n = 10$  for SHHα with mutant U1 snRNA,  $n = 27$  for SHHα with wild-type U1 snRNA,  $n = 23$  for SHHβ with wild-type U1 snRNA,  $n = 24$  for SHHγ with wild-type U1 snRNA,  $n = 46$  for SHHδ with mutant U1 snRNA).  $P$  values were determined using the two-sided log-rank test. +, censored case.



# Extended Data Fig. 5 | Intron-centric analysis of SHHδ medulloblastomas.

**a**, Quantification of alternative splicing events by SHH subtype, as detected by intron-centric alternative splicing analysis ( $n = 30$  of each subtype). Bar plot shows adjusted standardized residual of included alternative splicing events. Positive values indicate a relatively higher number, and negative values indicate a relatively lower number among subtypes. **b**, Volcano plots of alternative

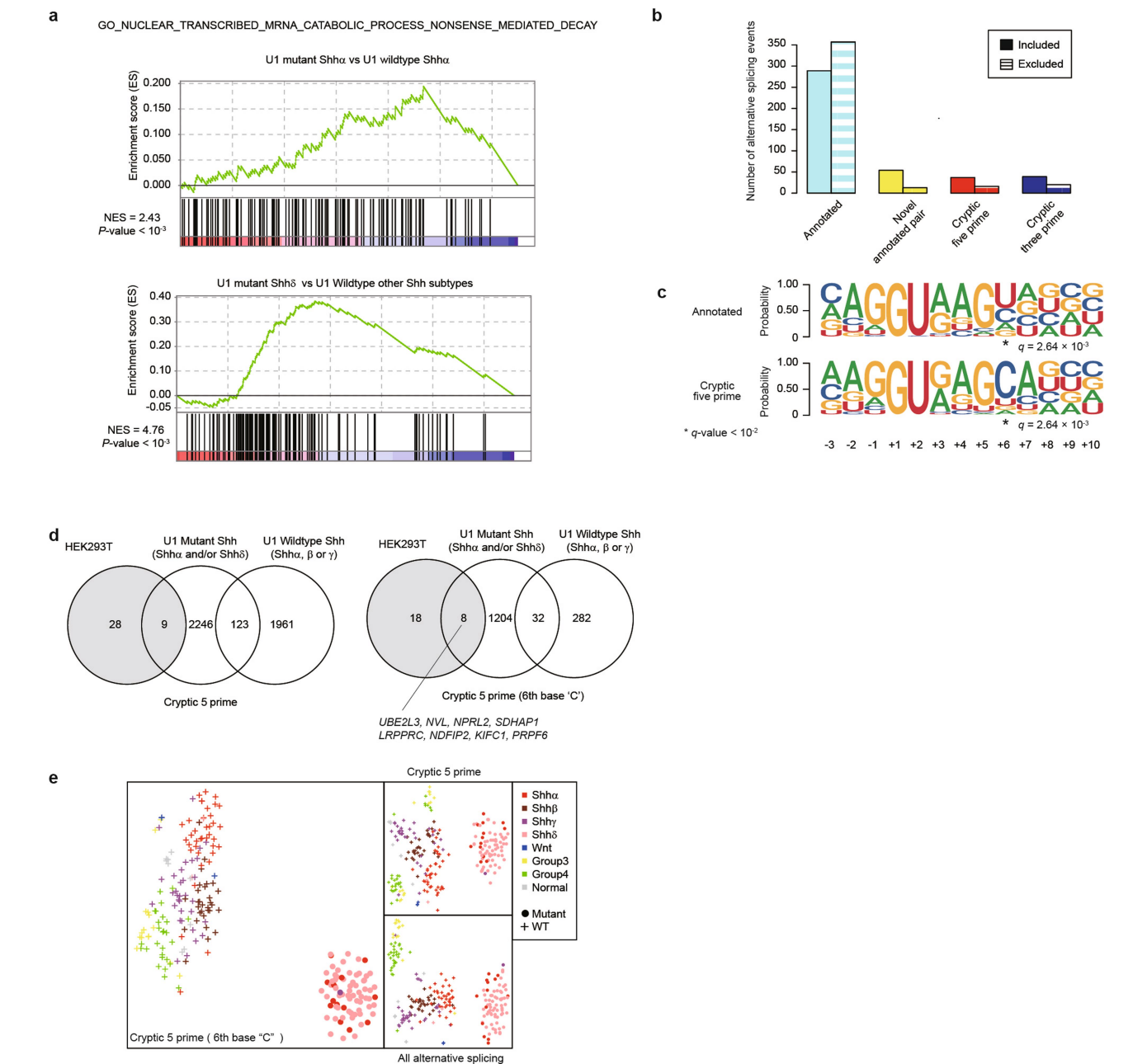
splicing events ( $n = 30$  of each subtype). Significant events ( $FDR < 0.01$  and absolute log effect size  $> 1.5$  calculated using LeafCutter (Methods)) are illustrated by colour. Alternative splicing events of *PTCH1* and *GLI2* with the highest effect size are annotated. **c**, Splice-site sequences of included alternative splicing events by subtype ( $n = 30$  of each subtype). Asterisk denotes nucleotide sites with  $q$  value  $< 10^{-2}$  ( $\chi^2$  test and Benjamini-Hochberg method).



Extended Data Fig. 6 | See next page for caption.

**Extended Data Fig. 6 | Intron-centric analysis of SHHα medulloblastomas.** **a, b,** Quantification (**a**) and proportion (**b**) of alternative splicing events between SHHα medulloblastoma with mutant U1 snRNA ( $n = 13$ ), and SHHα medulloblastoma with wild-type U1 snRNA ( $n = 39$ ) as detected by intron-centric alternative splicing analysis.  $P$  value was calculated by  $\chi^2$  test. **c,** Volcano plots of alternative splicing events ( $n = 13$  for SHHα with mutant U1 snRNA,  $n = 39$  for SHHα with wild-type U1 snRNA). The x axis shows the difference of per cent spliced in calculated using LeafCutter. Significant events ( $FDR < 0.01$  and absolute log effect size  $> 1.5$ , calculated by LeafCutter (Methods)) are illustrated by colour. **d,** Splice-site sequences of included alternative splicing events in

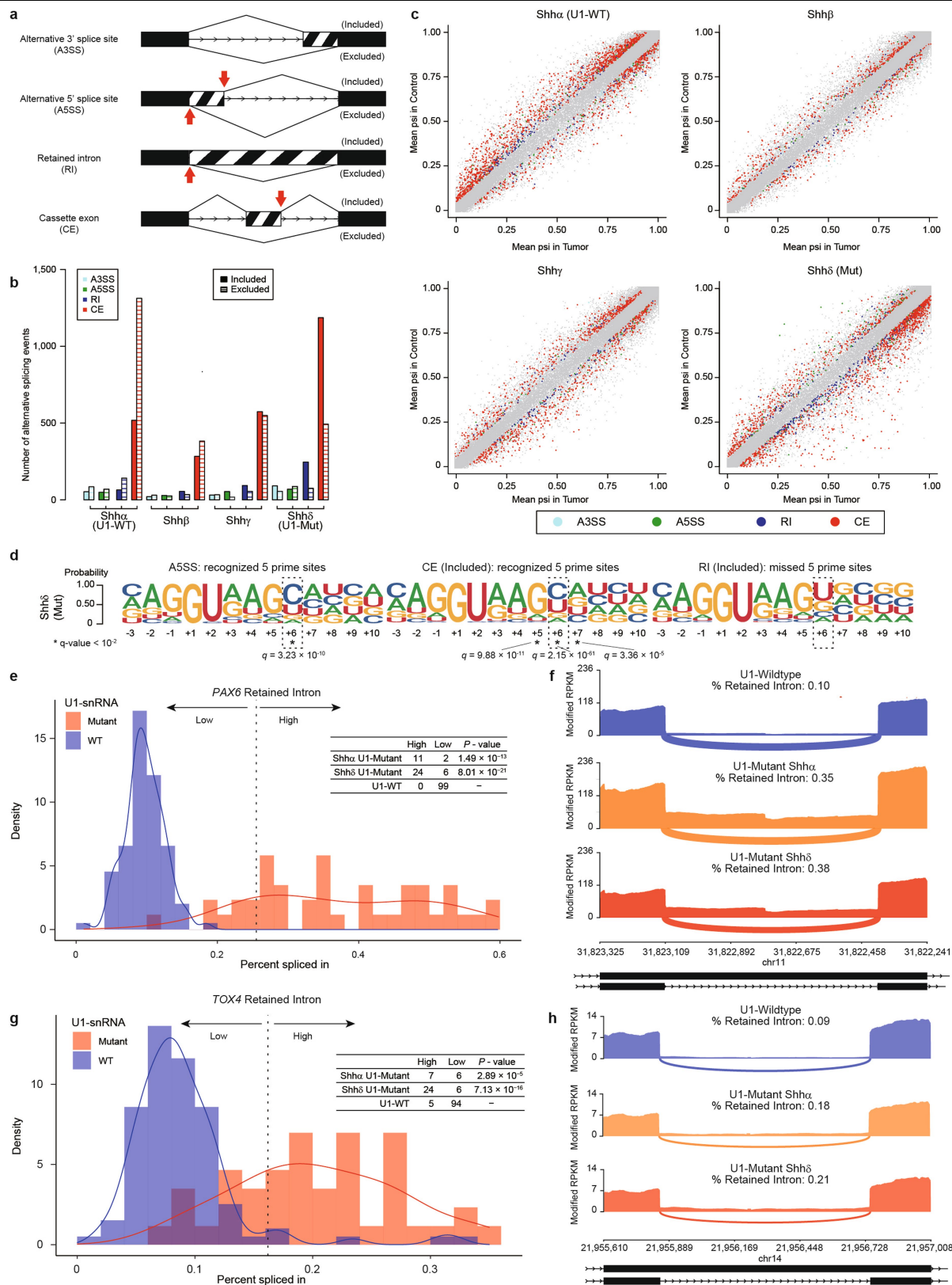
SHHα with mutant U1 snRNA ( $n = 13$  for SHHα with mutant U1 snRNA,  $n = 39$  for SHHα with wild-type U1 snRNA). Size and colour for each circle indicate the  $q$  values and Cramer's  $V$  values for each nucleotide position ( $q$  values were calculated by  $\chi^2$  test and Benjamini-Hochberg method; the precise values are given in Supplementary Table 11). **e,** Residual analysis of 5' splice-site sequences of annotated and cryptic 5' alternative splicing ( $n = 13$  for SHHα with mutant U1 snRNA,  $n = 39$  for SHHα with wild-type U1 snRNA). The size and colour of each circle denote the two-sided  $P$  value, and adjusted standardized residual calculated by Haberman's method. The precise values are given in Supplementary Table 11.



**Extended Data Fig. 7 | Nonsense-mediated decay pathway in SHH medulloblastoma with mutant U1 snRNA, and exogenous expression analyses.** **a**, Enrichment plots of 'GO nuclear transcribed mRNA catabolic process nonsense mediated decay' by GSEA between SHH $\delta$  with mutant U1 snRNA ( $n = 30$ ) and other subtypes of SHH medulloblastoma with wild-type U1 snRNA ( $n = 90$ ), and SHH $\alpha$  with mutant U1 snRNA ( $n = 13$ ) and SHH $\alpha$  with wild-type U1 snRNA ( $n = 39$ ).  $P$  values were calculated using gsea v.3.0 (Methods). **b**, Quantification of alternative splicing events between HEK-293T cells with mutant U1 snRNA and HEK-293T cells with wild-type U1 snRNA, as detected by intron-centric alternative splicing analysis. **c**, Splice-site sequences of included alternative splicing events in HEK-293T cells with mutant U1 snRNA. Asterisk denotes nucleotide sites with  $q$  value <  $10^{-2}$  ( $\chi^2$  and Benjamini-Hochberg

method). **d**, Comparison of the extent of overlap between detected alternative splicing events by SHH medulloblastoma (either of SHH $\alpha$  or of SHH $\delta$ ) with mutant U1 snRNA, SHH medulloblastoma (either of SHH $\alpha$ , SHH $\beta$  or SHH $\gamma$ ) with wild-type U1 snRNA and HEK-293T cells with mutant U1 snRNA exogenous expression. Left, alternatively spliced events with cryptic 5' sites. Right, alternatively spliced events with cryptic 5' sites and C base at the 6th intron. **e**, Alternative splicing signatures by  $t$ -SNE analysis. Left, the per cent spliced-in values of detected cryptic 5' alternative splicing events, with a 'C' nucleotide at the 6th base in the intron from the 5' splice site. Top right, per cent spliced-in values of all cryptic 5' alternative-splicing events. Bottom right, per cent spliced-in values of all alternative splicing events.

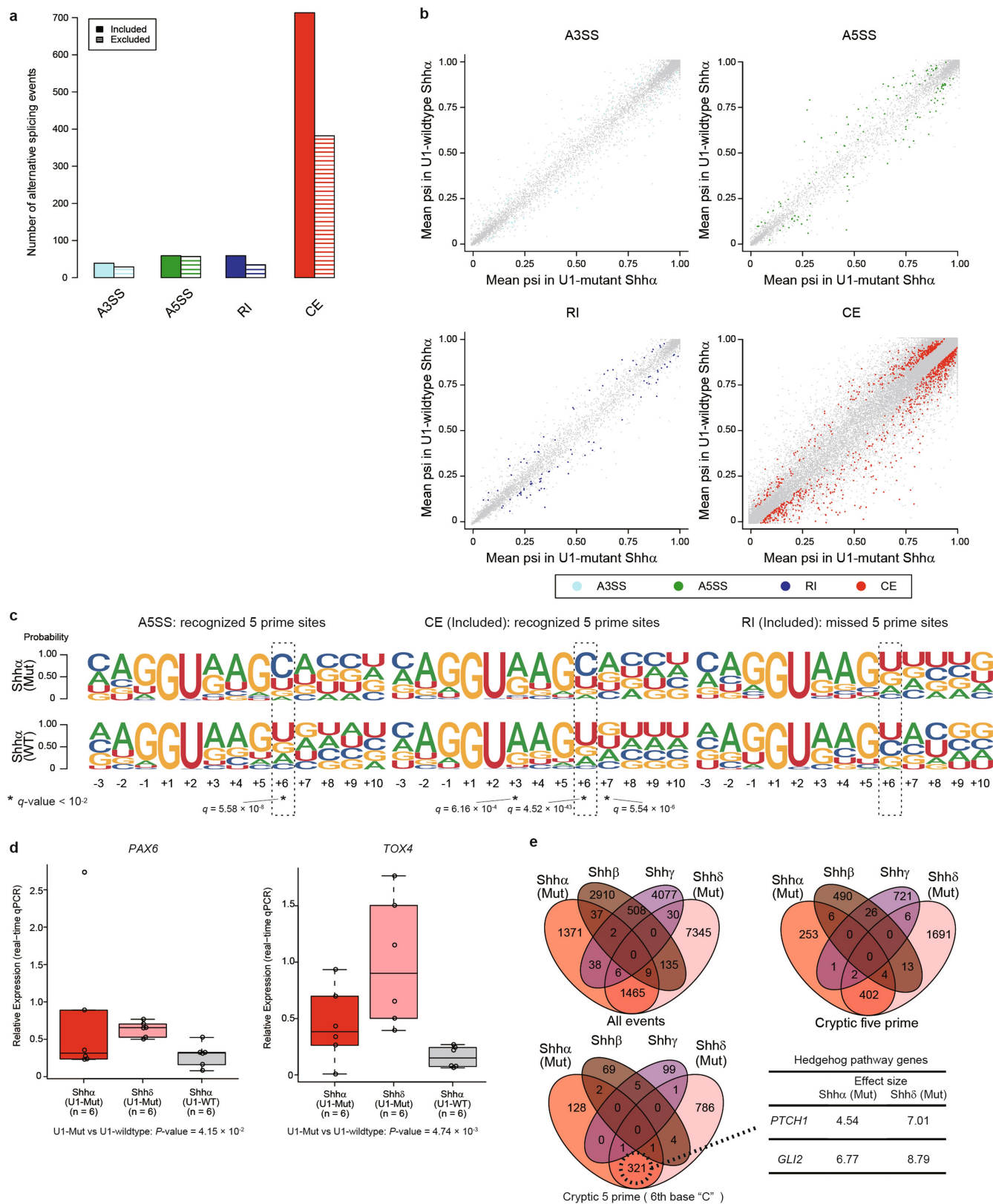




**Extended Data Fig. 8** | See next page for caption.

**Extended Data Fig. 8 | Retained introns inactivate tumour-suppressor genes in tumours with U1 snRNA r.3A>G mutation.** **a**, Illustration of the different types of alternative splicing events analysed using rMATS ( $n = 30$  of each subtype). Red arrows indicate expected 5' prime sites recognized by the mutant U1 snRNA. **b**, Quantification of alternative splicing events by subtype of SHH medulloblastoma, as detected by exon-centric alternative splicing analysis. **c**, Scatter plots of alternative splicing events ( $n = 30$  of each subtype). The x axis shows the difference of per cent spliced in (psi) calculated using rMATS. Different types of significant event ( $FDR < 0.01$  and absolute differential psi  $> 0.05$  calculated using rMATS (Methods)) are illustrated by different colours, as annotated. **d**, Splice-site sequences of alternative 5' splice site (A5SS), included cassette exon (CE) and included retained intron (RI) events in SHH $\delta$  with mutant U1 snRNA ( $n = 30$ ). Each event corresponds to a red arrow cartoon in **a**. Asterisk denotes nucleotide sites with  $q$  value  $< 10^{-2}$  ( $\chi^2$  test and Benjamini–Hochberg method). **e**, Distribution of per cent spliced in for *PAX6* based on U1 snRNA mutation status ( $n = 13$  for SHH $\alpha$  with mutant U1 snRNA,  $n = 30$  for SHH $\delta$  with mutant U1 snRNA,  $n = 99$  for SHH medulloblastoma with wild-type U1 snRNA ( $n = 90$ ) and normal brain tissue ( $n = 9$ )). Dashed line defines threshold that divides the dataset into two groups ( $k$ -means method). The table displays the number of samples above the threshold (high) or below (low) based on

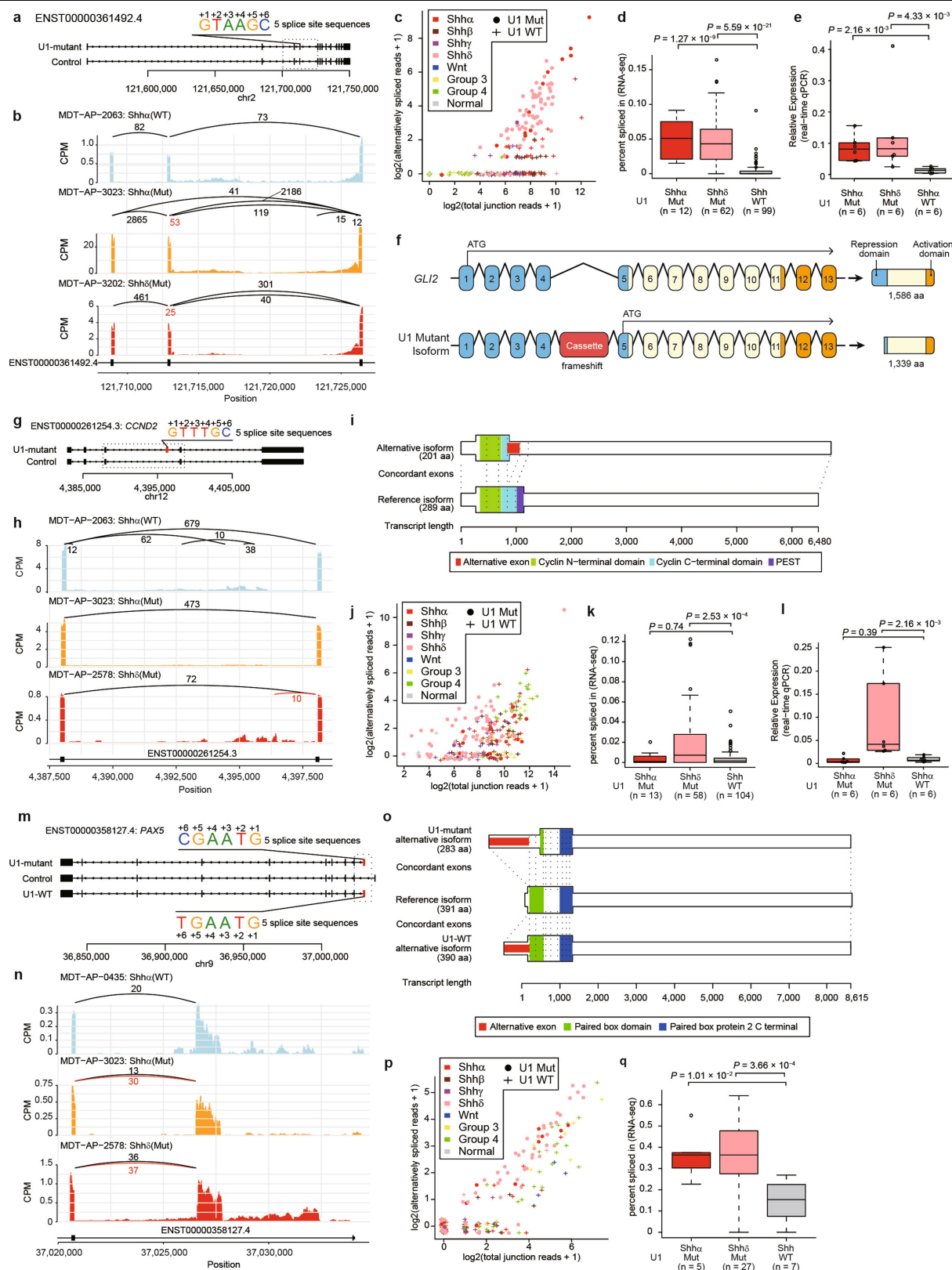
mutational status.  $P$  value was calculated using two-sided Fisher's exact test compared to samples with wild-type U1 snRNA. Samples with mutant U1 snRNA are indicated in pink, and wild-type samples in blue. **f**, Sashimi plot of splicing of *PAX6* based on mutational status determined by exon-centric alternative splicing analysis (rMATS). The bar plot shows modified FPKM. Numbers refer to average junctional reads across all samples. Annotated exon tracks are shown below, with genomic positions marked. **g**, Distribution of per cent spliced in for *TOX4* based on U1 snRNA mutation status ( $n = 13$  for SHH $\alpha$  with mutant U1 snRNA,  $n = 30$  for SHH $\delta$  with mutant U1 snRNA,  $n = 99$  for SHH medulloblastoma with wild-type U1 snRNA ( $n = 90$ ) and normal brain tissue ( $n = 9$ )). Dashed line defines threshold that divides the dataset into two groups ( $k$ -means method). The table displays the number of samples above the threshold (high) or below (low) based on mutational status.  $P$  value was calculated using two-sided Fisher's exact test compared to samples with wild-type U1 snRNA. Samples with mutant U1 snRNA are indicated in pink, and wild-type samples in blue. **h**, Sashimi plot of splicing of *TOX4* based on mutational status determined by exon-centric alternative splicing analysis (rMATS). The bar plot shows modified FPKM. Numbers refer to average junctional reads across all samples. Annotated exon tracks are shown below, with genomic positions marked.



**Extended Data Fig. 9** | See next page for caption.

**Extended Data Fig. 9 | Exon-centric analysis of SHHα medulloblastomas and overlapped splicing events.** **a**, Quantification of alternative splicing events between SHHα medulloblastoma with mutant U1 snRNA and SHHα medulloblastoma with wild-type U1 snRNA, as detected by exon-centric alternative splicing analysis. **b**, Scatter plots of alternative splicing events ( $n = 13$  for SHHα with mutant U1 snRNA,  $n = 39$  for SHHα with wild-type U1 snRNA). The  $x$  axis shows the difference of per cent spliced in calculated using rMATS. Different types of significant event (FDR < 0.01 and absolute differential psi > 0.05 calculated using rMATS (Methods)) are illustrated by different colours, as annotated. **c**, Splice-site sequences of alternative 5' splice sites, included cassette exon and included retained intron events in SHHα medulloblastoma with mutant U1 snRNA and SHHα medulloblastoma with wild-type U1 snRNA.

Each event corresponds to a red arrow cartoon in Extended Data Fig. 8a. Asterisk denotes nucleotide sites with  $q$  value <  $10^{-2}$  ( $\chi^2$  test and Benjamini–Hochberg method). **d**, Box plot of fold changes in expression of the alternatively spliced isoform as compared to the wild-type isoform in subsets of SHH medulloblastoma, as determined by real-time qPCR. In the box plots, the centre lines show data median; box limits indicate the IQR from the 25th and 75th percentiles; lower and upper whiskers extend 1.5× the IQR. Outliers are represented by individual points.  $P$  values were calculated using two-sided Wilcoxon rank-sum test. **e**, Comparison of the extent of overlap between splicing events by subtype of SHH medulloblastoma and U1 snRNA mutational status. Effect sizes are calculated by LeafCutter with an absolute effect-size threshold of 1.5.



**Extended Data Fig. 10** | See next page for caption.



## Extended Data Fig. 10 | Aberrant splicing of oncogenes and tumour-suppressor genes in SHH medulloblastoma with mutant U1 snRNA.

**a**, Overview of cryptic alternative splicing of *GLI2*, demonstrating the position of a cryptic cassette exon with the 5' splice-site sequence. **b**, Sashimi plot of splicing of *GLI2* in representative cases. The bar plot shows counts per million reads. Numbers are of junctional reads; reads for the mutant U1 snRNA isoform are in red. **c**, Scatter plot comparing detected alternatively spliced read and total junction reads that share a 3' splice site. Jittering was performed for both values. **d**, Per cent spliced in values for SHHα with mutant U1 snRNA, SHHδ with mutant U1 snRNA and SHH medulloblastoma with wild-type U1 snRNA (all subtypes of SHH medulloblastoma). **e**, Box plot of fold changes in expression of the alternatively spliced isoform as compared to the wild-type isoform of *GLI2* in subsets of SHH medulloblastoma, determined by real-time qPCR. **f**, Illustration of canonical and cryptic isoforms of *GLI2*. Translation start sites are indicated by an ATG arrow. Resulting proteins (and sizes) are displayed for each isoform. Repression and activation domains are indicated in blue and orange, respectively. **g**, Overview of cryptic alternative splicing of *CCND2*, illustrating the position of a cryptic cassette exon with the 5' splice-site sequence. **h**, Sashimi plot of representative cases demonstrates alternative splicing at the *CCND2* locus. Numbers illustrate junctional reads. Junctional reads specific to U1 snRNA mutants are in red. **i**, The canonical isoform and the cryptic isoform of *CCND2*.

**j**, Scatter plot comparing detected alternatively spliced read and total junction reads that shared a 3' splice site. Jittering was performed for both values. **k**, Per cent spliced in values for U1-mutant SHHα ( $n = 13$ ), U1-mutant SHHδ ( $n = 58$ ), and U1-wildtype SHH (all SHH subtypes,  $n = 104$ ). **l**, Real-time qPCR comparing the expression of the cryptic isoform of *CCND2* demonstrates high levels of expression of *CCND2* restricted to SHHδ cases ( $n = 6$  for SHHα with mutant U1 snRNA,  $n = 6$  for SHHδ with mutant U1 snRNA,  $n = 6$  for SHHα with wild-type U1 snRNA). **m**, Overview of cryptic alternative splicing of *PAX5*, illustrating the position of a cryptic cassette exon with the 5' splice-site sequence. **n**, Sashimi plot of representative cases demonstrates alternative splicing at the *PAX5* locus. Numbers denote junctional reads. Junctional reads specific to U1 snRNA mutants are in red. **o**, The canonical isoform and the cryptic isoform of *PAX5*. **p**, Scatter plot comparing detected alternatively spliced read and total junction reads that shared a 3' splice site. Jittering was performed for both values. **q**, Per cent spliced in values by SHHα with mutant U1 snRNA ( $n = 5$ ), SHHδ with mutant U1 snRNA ( $n = 27$ ) and SHH medulloblastoma with wild-type U1 snRNA (all subtypes of SHH medulloblastoma,  $n = 7$ ). In all box plots, centre lines show data median; box limits indicate the IQR from the 25th and 75th percentiles; lower and upper whiskers extend 1.5× the IQR. Outliers are represented by individual points. *P* values were calculated using two-sided Wilcoxon rank-sum tests.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used to collect data.

Data analysis

Burrows-Wheeler Aligner (v0.7.8), biobambam (v0.0.148), GATK (v3.5.0 and v3.8.0), EBCall (v0.2.1), Varscan2 (v2.4.3), Strelka (v1.0.15), SomaticSniper (v1.0.5.0), Viramid (v1.1.0), Platypus (v0.8.1), Seurat (v2.5), Control-FREEC (v10.3), STAR (v2.5.1b), LeafCutter (v0.2.7), rMATS (v4.0.1), MISO (v0.5.4), htseq (v0.6.0), gsea (v3.0), Picard tool (v2.18.0), bowtie (v2.3.4.1), Affymetrix Power Tools (v1.18.2), PennCNV (v1.0.3), ASCAT (v2.4.3), GISTIC (v2.0.23), R (v3.3.0), R packages - Rtsne (v0.13), ggseqlogo (v0.157), survival (v2.40.1), Epi (v2.30).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data have been deposited in the European Genome-Phenome Archive (EGA) and Gene Expression Omnibus (GEO) : RNA-seq (EGAD00001001899, and EGAD00001004958), whole genome sequence (EGAD00001003125 and EGAD00001004347) and RNA-seq of exogenous expression analyses (GSE128005).

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences    ☐ Behavioural & social sciences    ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined by the availability of the human samples. We used all sequencing data we have.
Data exclusions	In ICGC medulloblastoma data, we excluded three samples. The reasons are a truncated file, obvious tumor contamination in the matched control sample, and single-end sequencing data. The exclusion criteria was not pre-established. These three files cannot be analyzed with our mutation call pipeline described in Methods section.
Replication	rhAmp SNP genotyping were done three times. All attempts at replication were successful. All qPCR was done in technical triplicate aside from biological replicates.
Randomization	For alternative splicing analysis, we choose samples randomly using sample function in R software.
Blinding	Blinding was not relevant for our study since this is an exploratory study and blinding is impossible or unlikely to affect the results or interpretation of the results.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	HEK293T was obtained from American Type Culture Collection (ATCC)
Authentication	HEK293T cell was not authenticated.
Mycoplasma contamination	Mycoplasma negative
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	No commonly misidentified cell lines were used in this study.

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Samples were collected at diagnosis after informed consent was obtained from subjects as part of the Medulloblastoma Advanced Genomics International Consortium. Research participants were patients with confirmed diagnosis of medulloblastoma at local centers. All cases used in this study were primary cases.
Recruitment	Patients were recruited by investigators from each local center. There is no bias of recruitment because patients are not prescreened. Potential self-selection bias or other biases were not identified.

Ethics oversight

The Hospital for Sick Children

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# The U1 spliceosomal RNA is recurrently mutated in multiple cancers

<https://doi.org/10.1038/s41586-019-1651-z>

Received: 3 September 2018

Accepted: 3 September 2019

Published online: 9 October 2019

Shimin Shuai<sup>1,2,13</sup>, Hiromichi Suzuki<sup>3,4,13</sup>, Ander Diaz-Navarro<sup>5,6,13</sup>, Ferran Nadeu<sup>5,7</sup>, Sachin A. Kumar<sup>3,4,8</sup>, Ana Gutierrez-Fernandez<sup>5,6</sup>, Julio Delgado<sup>5,9</sup>, Magda Pinyol<sup>5,10</sup>, Carlos López-Otín<sup>5,6</sup>, Xose S. Puente<sup>5,6</sup>, Michael D. Taylor<sup>3,4,11</sup>, Elías Campo<sup>5,7,12</sup> & Lincoln D. Stein<sup>1,2\*</sup>

Cancers are caused by genomic alterations known as drivers. Hundreds of drivers in coding genes are known but, to date, only a handful of noncoding drivers have been discovered—despite intensive searching<sup>1,2</sup>. Attention has recently shifted to the role of altered RNA splicing in cancer; driver mutations that lead to transcriptome-wide aberrant splicing have been identified in multiple types of cancer, although these mutations have only been found in protein-coding splicing factors such as splicing factor 3b subunit 1 (*SF3B1*)<sup>3–6</sup>. By contrast, cancer-related alterations in the noncoding component of the spliceosome—a series of small nuclear RNAs (snRNAs)—have barely been studied, owing to the combined challenges of characterizing noncoding cancer drivers and the repetitive nature of snRNA genes<sup>1,7,8</sup>. Here we report a highly recurrent A>C somatic mutation at the third base of U1 snRNA in several types of tumour. The primary function of U1 snRNA is to recognize the 5′ splice site via base-pairing. This mutation changes the preferential A–U base-pairing between U1 snRNA and the 5′ splice site to C–G base-pairing, and thus creates novel splice junctions and alters the splicing pattern of multiple genes—including known drivers of cancer. Clinically, the A>C mutation is associated with heavy alcohol use in patients with hepatocellular carcinoma, and with the aggressive subtype of chronic lymphocytic leukaemia with unmutated immunoglobulin heavy-chain variable regions. The mutation in U1 snRNA also independently confers an adverse prognosis to patients with chronic lymphocytic leukaemia. Our study demonstrates a noncoding driver in spliceosomal RNAs, reveals a mechanism of aberrant splicing in cancer and may represent a new target for treatment. Our findings also suggest that driver discovery should be extended to a wider range of genomic regions.

To determine the extent of U1 snRNA (hereafter, U1) mutations in cancer, we first screened 2,583 whole-genome sequenced donors across 37 tumour types from the ‘Pan-Cancer Analysis of Whole Genomes’ (PCAWG) project<sup>9</sup> (Supplementary Table 1). The human genome (GRCh37) has 7 genes with the same canonical 164-bp U1 sequence, and more than 130 pseudogenes with variant U1 sequences; the flanking sequences of the genes and pseudogenes are also highly similar<sup>7,8</sup> (Supplementary Note). We therefore called somatic mutations across canonical U1 genes by using reads mapped only to them (Extended Data Fig. 1a), and reported all possible mutated genes for each U1 mutation. Within the 2,434 donors who had sufficient coverage, we identified 277 somatic mutations in U1 genes that affected 240 donors, across 30 tumour types

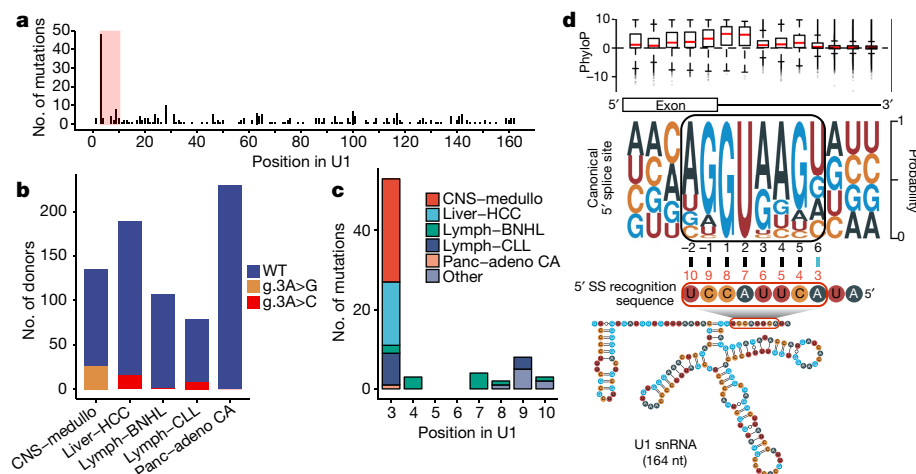
(Fig. 1a, Supplementary Table 2). These mutations spanned 100 of the 164 bases of U1, but only 2 positions (base 3 and base 28) were mutated in more than 5% of donors in at least 1 tumour type.

Base 28 of U1 falls in a stem loop, and was recurrently mutated in 4 out of 23 (17.4%) bladder cancers. This may be related to the previously reported higher mutation rate among palindrome loops in bladder cancer<sup>10</sup>. The third base of U1 contained 27 A>G and 21 A>C mutations across five types of tumour. This base forms part of the highly conserved 5′ splice-site recognition sequence (nucleotides 3–10) of U1, which base-pairs directly with 5′ splice site<sup>11</sup>. Five additional A>C mutations were recovered from samples with insufficient coverage (Supplementary Note). Collectively (Fig. 1b), the A>G mutation was found in 26 out of

<sup>1</sup>Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. <sup>2</sup>Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada.

<sup>3</sup>Developmental and Stem Cell Biology Program, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>4</sup>The Arthur and Sonia Labatt Brain Tumour Research Centre, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>5</sup>Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain. <sup>6</sup>Departamento de Bioquímica y Biología Molecular, Instituto Universitario de Oncología (IUOPA), ISPA, Universidad de Oviedo, Oviedo, Spain. <sup>7</sup>Patología Molecular de Neoplasias Limfoides, Institut d’Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. <sup>8</sup>Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. <sup>9</sup>Servei Hematologia, Hospital Clínic de Barcelona, Barcelona, Spain. <sup>10</sup>Unitat de Genòmica, Institut d’Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. <sup>11</sup>Division of Neurosurgery, The Hospital for Sick Children, Toronto, Ontario, Canada. <sup>12</sup>Unitat Hematopatologia, Hospital Clínic de Barcelona, Universitat de Barcelona, Barcelona, Spain. <sup>13</sup>These authors contributed equally: Shimin Shuai, Hiromichi Suzuki, Ander Diaz-Navarro. \*e-mail: [lincoln.stein@gmail.com](mailto:lincoln.stein@gmail.com)





**Fig. 1 | Overview of somatic mutations in U1.** **a**, Distribution of mutations in canonical U1 genes. The red shaded region indicates the 5' splice-site recognition sequence. **b**, Recurrent mutations at the third base of U1 across five tumour types. Adeno CA, adenocarcinoma; CNS, central nervous system; BNHL, B cell non-Hodgkin lymphoma; medullo, medulloblastoma; panc, pancreas. **c**, Distribution of mutations within the 5' splice-site recognition sequence by

tumour type. **d**, The RNA-RNA interaction between U1 and the 5' splice site. Bases 3 to 10 of U1 (red box and numbering) can base-pair with the 5' splice site (black box and numbering). The base-pairing affected by the g.3A>C mutation is in blue. PhyloP scores show the conservation levels of human canonical 5' splice site (n = 344,580 introns). Centre line, median; box limits, upper and lower quartiles; whiskers, 1.5 × interquartile range; points, outliers.

135 (19.3%) cases of medulloblastoma (which are described in detail in the accompanying paper<sup>12</sup>) and 1 out of 230 (0.4%) cases of pancreatic adenocarcinomas. The A>C mutation (hereafter g.3A>C) was found in 8 out of 78 (10.3%) cases of chronic lymphocytic leukaemia (CLL), 16 out of 189 (8.5%) cases of hepatocellular carcinoma (HCC) and 2 out of 107 (1.9%) cases of B cell non-Hodgkin lymphomas. Mutations were also found in other bases of the 5' splice-site recognition sequence, but at a much lower frequency (Fig. 1c); a preponderance of non-third-base mutations (9 out of 20) was observed in B cell non-Hodgkin lymphomas.

To enlarge our sample size, we used widespread alterations in splicing and expression patterns of samples with U1 mutations to infer the mutational status of the third base in tumours that were associated with RNA-sequencing (RNA-seq) data only (Extended Data Fig. 1b, c and 2f). Using machine learning (Methods, Supplementary Note), we predicted that an additional 4 out of 240 (1.7%) cases of CLL and 14 out of 321 (4.4%) cases of HCC contain the g.3A>C mutation. To benchmark the classifier, we validated the g.3A>C status for 298 cases of CLL using a PCR-based SNP genotyping system (rhAmp) (Methods). Only one sample showed an inconsistent genotype, which was treated as mutated in subsequent analysis (Supplementary Note). By combining the results based on whole-genome sequencing and analyses of the transcriptome, the recurrent g.3A>C mutation was found in 12 out of 318 (3.8%) donors with CLL and 30 out of 510 (5.9%) donors with HCC (Extended Data Fig. 1d, e, Supplementary Table 3).

The splice-site consensus motif prefers a U at the sixth position of the 5' splice site; we hypothesized that the g.3A>C mutation could shift this preference towards a G (which we hereafter term the G6 5' splice site) (Fig. 1d). Using RNA-seq data from matched cases of CLL (11 with U1 mutation versus 254 with wild-type U1) and HCC (20 with U1 mutation versus 367 with wild-type U1), we performed differential splicing analysis using annotation-free and intron-centric software (LeafCutter)<sup>13</sup> (Methods). We identified 3,193 and 533 differentially spliced introns in 1,519 and 303 genes (LeafCutter  $q < 0.1$  and absolute  $\log_2$ (effective size) > 1) in CLL and HCC, respectively (Fig. 2a, Supplementary Table 4).

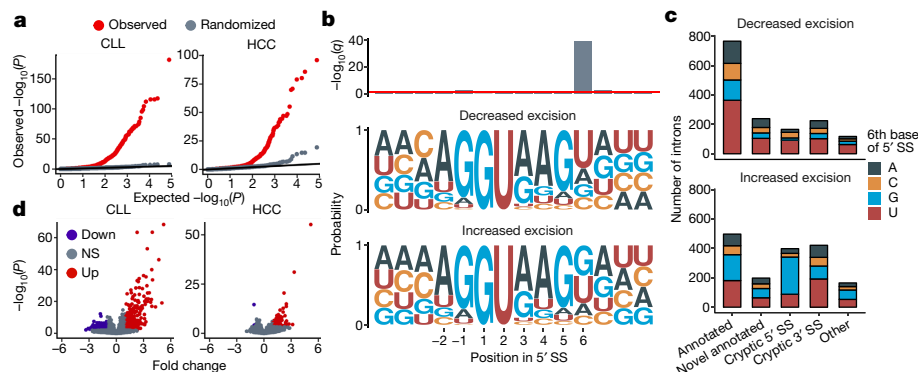
For each intron, we further determined its direction of change using the change in per cent spliced in ( $\Delta$ PSI) (Extended Data Fig. 2b). When comparing the base composition of the 5' splice site among introns with increased excision ( $\Delta$ PSI > 0) and decreased excision ( $\Delta$ PSI < 0) in samples with U1 mutation, we observed significant differences at the sixth position for both CLL ( $\chi^2$  test,  $P = 3.3 \times 10^{-41}$ ) (Fig. 2b) and HCC ( $\chi^2$

test,  $P = 6.7 \times 10^{-8}$ ) (Extended Data Fig. 2c). Consistent with expectations, introns with increased excision were highly enriched in the G6 5' splice site relative to introns with decreased excision (38.4% versus 16.4% in CLL; 47.0% versus 22.1% for HCC) (Fig. 2b), or to genome-wide canonical introns (18.9%) (Fig. 1d). In consequence, we observed many novel splicing events in samples with U1 mutation—especially for splicing with the cryptic G6 5' splice site in both types of tumour (Fig. 2c, Extended Data Fig. 2d). Together, these data support the hypothesis that the g.3A>C mutation increases the splicing rate of the G6 5' splice site.

Because splicing and expression frequently correlate<sup>5,14</sup>, we also conducted differential expression analysis for the g.3A>C mutation (Extended Data Fig. 2e). This analysis revealed 869 and 68 differentially expressed genes ( $q < 0.1$  and absolute  $\log_2$ -transformed fold change > 1) (Supplementary Table 4) for CLL and HCC, respectively. More genes were upregulated than downregulated in the samples with U1 mutations: 561 out of 869 and 66 out of 68 in cases of CLL and HCC, respectively (Fig. 2d).

We next investigated genes affected by the U1 mutation. We found that 84 and 16 genes in the Cancer Gene Census (v.84) were mis-spliced in cases of CLL and HCC, respectively<sup>15</sup>. Among these genes, 44 and 10—respectively—had increased excision of G6 5' splice site introns, including known drivers of CLL and HCC (Supplementary Table 4). The most significant mis-spliced cancer gene in CLL was musashi RNA binding protein 2 (*MSI2*) (LeafCutter  $q = 1.2 \times 10^{-112}$ ); CLL with U1 mutations exclusively expressed a cryptic exon that contains a premature termination codon, and was associated with a G6 5' splice site (Fig. 3a, Extended Data Fig. 3a). A similar pattern was observed for the gene DNA polymerase delta 1, catalytic subunit (*POLD1*) ( $q = 4.2 \times 10^{-33}$ ) (Fig. 3b, Extended Data Fig. 3a). As the cryptic exon affected the polymerase—but not the exonuclease—domain of *POLD1*, the g.3A>C mutation was not associated with a higher mutation burden.

We also found mis-splicing in other genes related to CLL biology, such as the hyaluronic acid receptor gene CD44 molecule (Indian blood group) (*CD44*). *CD44* was the most significantly differentially spliced gene (LeafCutter  $q = 5.1 \times 10^{-178}$ ). Alternative splicing of *CD44* is tissue-specific and has previously been associated with processes such as lymphocyte homing and tumorigenesis; the gene is also thought to regulate anti-apoptosis signalling in CLL<sup>16,17</sup>. Patients with wild-type CLL expressed predominantly the standard isoform (CD44s, which does not contain exon v2–v10) (Fig. 3c), whereas cases of CLL with U1 mutations overexpressed multiple variant isoforms (CD44v)—presumably



**Fig. 2 | Global gene splicing and expression changes associated with the g.3A>C mutation. a**,  $P$  value quantile–quantile plots for differential splicing analysis.  $P$  values are from LeafCutter. **b**, 5' splice site (SS) for introns with increased ( $n=1,657$  introns) or decreased excision ( $n=1,536$  introns) in cases of CLL with U1 mutation ( $n=11$  patients). The top bar chart shows  $q$  values from  $\chi^2$  tests for base composition difference; the red line indicates  $q=0.1$ . **c**, Category of mis-splicing events in CLL. Extended Data Figure 2a provides the definitions

of each category. The number of introns is coloured by the sixth base of 5' splice site. **d**, Volcano plots for differential expression analysis.  $P$  and  $q$  values are from limma. NS, not significant ( $q>0.1$  or  $\log_2$ -transformed fold change  $<1$ ). For **a**, **d**, biologically independent patient samples are used for CLL (11 with U1 mutation versus 254 with wild-type U1) and HCC (20 with U1 mutation versus 367 with wild-type U1).

because the presence of several G6 5' splice sites increased the excision rate of introns associated with variant exons (Fig. 3d). Another, similar example is ATP-binding cassette sub-family D member 3 (*ABCD3*), a fatty acid transporter for peroxisomes; two cryptic exons were expressed exclusively in cases of CLL with U1 mutations (Extended Data Fig. 3a–c). The consistent combination of frequent mis-splicing ( $q=7.1\times 10^{-76}$ ) and overexpression ( $\log_2$ -transformed fold change = 2.3 and  $q=3.7\times 10^{-60}$ ) in *ABCD3* enabled us to create a single-gene score that predicted g.3A>C mutational status with 100% accuracy in CLL (Extended Data Fig. 3d,

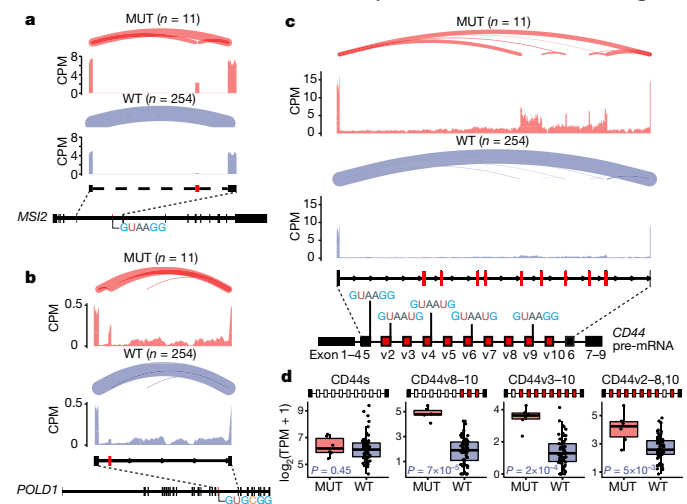
e). We experimentally validated the differentially spliced junctions of *MSI2*, *POLD1*, *CD44* and *ABCD3* using quantitative PCR (Extended Data Fig. 4a–c).

Genes the aberrant splicing of which introduces a premature termination codon are expected to be targeted by nonsense-mediated decay. However, mis-spliced forms of *MSI2*, *POLD1* and *ABCD3* that contained a premature termination codon were not downregulated in cases of CLL with U1 mutation, even though the distance between the premature termination codon and the final exon–exon junction exceeded the distance (55 nucleotides) needed for nonsense-mediated decay<sup>18</sup> (Extended Data Fig. 3a, b).

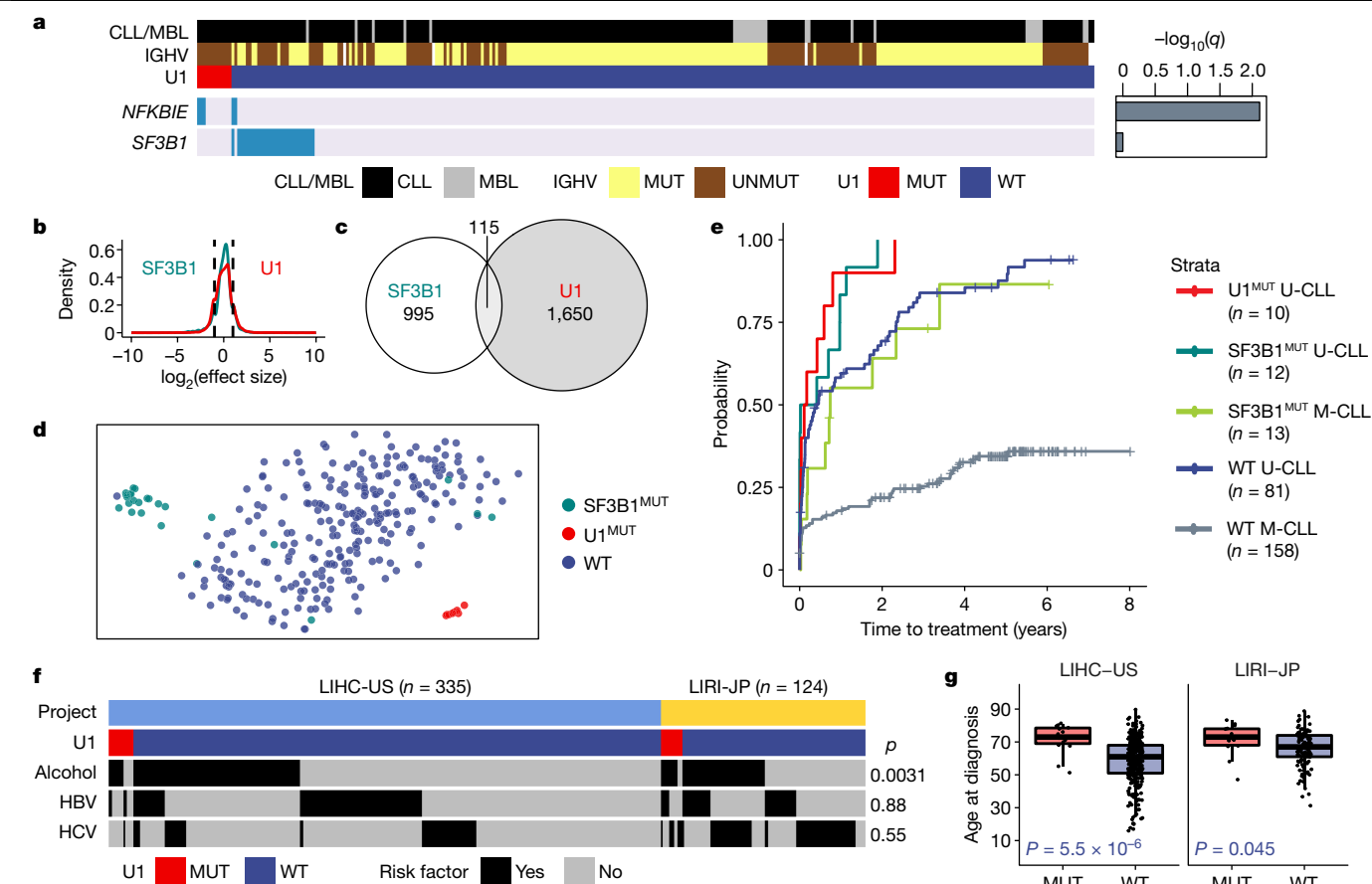
Next, we investigated the pathway-level changes that are associated with the g.3A>C mutation in CLL using gene-set enrichment analysis<sup>19</sup> (Supplementary Table 5). In CLL with U1 mutation, we found that genes related to mRNA transcription, RNA splicing, protein ubiquitination and telomere maintenance were upregulated (Extended Data Fig. 5a, b), whereas genes related to apoptosis, B cell receptor signalling and cytoplasmic ribosomes were downregulated (Extended Data Fig. 5c–g). The downregulation of ribosomal genes may explain the reduced rates of nonsense-mediated decay<sup>18</sup> that were noted earlier.

To validate our findings, we introduced exogenous U1 genes with or without the g.3A>C mutation into three CLL cell lines (JVM3, HG3 and MEC1). After confirming the exogenous expression of U1 (Extended Data Fig. 6a), we performed the same transcriptome analysis using cell-line RNA-seq data. In total, 7,238 introns in 2,365 genes were differentially spliced, and 459 genes were differentially expressed in cell lines that contained U1 mutations (Supplementary Table 4). Cell lines with U1 mutations also had many cryptic 5' splice-site splicing events, and more G6 5' splice site in introns with increased excision than in introns with decreased excision (34.4% versus 15.3%;  $\chi^2$  test  $P=1.9\times 10^{-75}$ ) (Extended Data Fig. 6b, c); 39.1% of the G6 5' splice site introns with increased excision in patients with U1 mutations were also shared by cell lines with U1 mutations (Extended Data Fig. 6d). In addition, cell lines with U1 mutations also had more genes upregulated (361 genes) than downregulated (97 genes) (Extended Data Fig. 6e), and shared many differentially expressed genes with primary CLL (Extended Data Fig. 6f, g). These data validate a causal link between the g.3A>C mutation and global splicing changes.

We further studied interactions between the g.3A>C mutation and other drivers of CLL and HCC. The g.3A>C mutation significantly co-occurred with the mutation of *NFKB1E* in CLL (Fisher's  $q=0.0077$ ) (Fig. 4a, Extended Data Fig. 7), the mutation of *APOB* in HCC (Mantel–Haenszel test  $q=0.018$ ) (Extended Data Fig. 9), and the mutation of the *TERT*



**Fig. 3 | Cancer-related genes that are mis-spliced in CLL with U1 mutations. a–c**, Sashimi plots showing mis-splicing patterns of *MSI2* (a), *POLD1* (b) and *CD44* (c). MUT ( $n=11$ ) and WT (wild type) ( $n=254$ ) represent samples of CLL with and without g.3A>C mutations, respectively. For each genotype (MUT in red; WT in blue), the three tracks from top to bottom show splice junctions, average expression levels in counts per million (CPM) and gene models, respectively. Each splice junction is shown as a curve and weighted by PSI values from LeafCutter. In gene models, black boxes are annotated exons, and red boxes indicate the cryptic (a, b) or variant (c) exons. The introns for *MSI2* are downscaled for better visualization (dashed lines). Gene models for *MSI2* (ENST00000284073.2) and *POLD1* (ENST00000599857.1) are based on primary transcripts; the gene model of *CD44* is shown as a cartoon. **d**, Isoform expression of *CD44*.  $P$  values are from two-sided Wilcoxon rank-sum tests. MUT ( $n=6$ ) and WT ( $n=61$ ) represents biologically independent CLL samples with and without g.3A>C mutations, respectively. For the box plot, centre line, box limits, whiskers and points represent the median, 25th and 75th percentiles,  $1.5\times$  interquartile range and individual samples, respectively.



**Fig. 4 | Driver alterations and clinical features related to the g.3A>C mutation.** **a**, Clinical features and selected driver events in CLL ( $n = 313$  patients). The right bar chart shows the Benjamin–Hochberg adjusted  $P$  values from two-sided Fisher’s exact tests. Complete driver events are in Extended Data Fig. 7. CLL/MBL, CLL or monoclonal B cell lymphocytosis. **b**, Distribution of effect size for mis-spliced introns in CLL with U1 mutations or *SF3B1* mutations. Dashed lines indicate the cutoff of absolute  $\log_2(\text{effect size}) = 1$ . **c**, Euler plot of mis-spliced intron clusters in CLL with U1 mutations or *SF3B1* mutations. **d**,  $t$ -distributed stochastic neighbour embedding plot showing CLL with U1 mutations or *SF3B1* mutations can be well-separated on the basis of mis-splicing patterns. **e**, Kaplan–Meier plot for time to first treatment in CLL. Patients with U1

or *SF3B1* mutations with U-CLL have a very similar disease course. **f**, Relationships between major risk factors and the g.3A>C mutation in cases of HCC.  $P$  values are from the Cochran–Mantel–Haenszel test (project code LIHC-US,  $n = 335$ ; project code LIRI-JP,  $n = 124$ ; definitions of project codes are provided in the Methods). HBV, infection with hepatitis B virus; HCV, infection with hepatitis C virus. **g**, Box plot for age at diagnosis in cases of HCC.  $P$  values are from two-sided Wilcoxon rank-sum tests. The LIHC-US has 15 patients with the g.3A>C mutation (MUT) and 336 without (WT); the LIRI-JP has 13 MUT and 116 WT. Centre line, median; box limits, 25th and 75th percentiles; whiskers, 1.5 $\times$  interquartile range; jitter points, individual samples.

promoter in one of two HCC projects (project code LIHC-US) (Fisher’s  $q = 0.016$ ). We found that none of the samples with U1 mutations had *SF3B1* mutations in CLL, although a larger dataset is needed for sufficient power to show mutual exclusion.

Because mutated *SF3B1* is also known to induce global splicing changes<sup>14</sup>, we compared samples with U1 mutation to samples with *SF3B1* mutations in CLL. Consistent with previous findings, *SF3B1* mutations induced many cryptic 3’ splice-site splicing events<sup>20</sup> (Extended Data Fig. 8a–c). Both mutations induced numerous mis-splicing events with small effect sizes (Fig. 4b), but tended not to share events (Fig. 4c, d). Using an exon-centric method, we found that CLL with U1 mutations tends to induce intron retention and suppress exon skipping, whereas CLL with *SF3B1* mutations demonstrated the opposite trend (Extended Data Fig. 8d, e). Notably, introns excised and exons retained in CLL with U1 mutations were enriched for the G6 5’ splice site (Extended Data Fig. 8f).

We next investigated the clinical relevance of the g.3A>C mutation. CLL has two major subtypes: one subtype in which the immunoglobulin heavy-chain variable regions (IGHV) are mutated (M-CLL), and a more aggressive subtype in which the IGHV are unmutated (U-CLL)<sup>21</sup>. The g.3A>C mutation was frequently found in cases of U-CLL (12 out of 105,

11.4%) (Fisher’s exact test  $P = 5.6 \times 10^{-6}$ ) (Fig. 4a) but not in M-CLL (0 out of 173). The mutation was also not found in monoclonal B cell lymphocytosis (MBL;  $n = 29$  cases), a lesion that precedes the overt leukaemia phase<sup>21</sup>. Moreover, the g.3A>C mutation in CLL was significantly associated with a shorter time to first treatment (log-rank test  $P = 1 \times 10^{-5}$ ), which indicates a more aggressive disease. The correlation with time to first treatment was significant even after adjusting for known prognostic markers, including disease stage (Binet stage), *SF3B1* mutations and IGHV status<sup>21,22</sup> (multivariate Cox model  $P = 0.043$ ) (Fig. 4e, Extended Data Fig. 10b). However, we observed no difference in overall survival between cases of CLL with U1 mutation and wild-type U1 (Extended Data Fig. 10a). We noted that 7 out of 10 cases of CLL with U1 mutation involved early-stage disease (Binet stage A) (Extended Data Fig. 10c), a hint that the mutation may appear at an early phase of the disease.

HCC has multiple risk factors, including infection with hepatitis B or C virus and heavy alcohol use<sup>23</sup>. We found that the U1 mutation was associated with increased alcohol intake (Mantel–Haenszel test,  $P = 0.0031$ ) but not with infection with hepatitis B or C virus (Fig. 4f). The mutation was also associated with increased age at diagnosis, but not with survival (Fig. 4g, Extended Data Fig. 10d–f). As in CLL, the mutation was also found in early disease stages of HCC (Extended Data Fig. 10g).

Here we provide an example of recurrent mutations in a noncoding splicing factor across multiple types of cancer. Splicing factor mutations in *SF3B1* and *SRSF2* that have previously been identified have been thought to promote tumorigenesis by inducing transcriptome-wide splicing changes that are subtle overall<sup>14,24</sup>. The same global effect is observed here for the U1 mutation. We also find mis-splicing of multiple known or putative cancer genes in cases of CLL and HCC with U1 mutations, which supports the theory that the tumorigenic effects of spliceosomal mutations are mediated by the production of specific aberrant isoforms<sup>25</sup>, although detailed functional analysis is required to confirm the role of these isoforms.

The U1 mutation has potential clinical applications. Besides its use as an independent prognostic marker in CLL, the mutation may also represent an opportunity for treatment. Inhibitors of *SF3B1* have previously been demonstrated to preferentially kill tumour cells that contain splicing factor mutations via synthetic lethality<sup>26,27</sup>; this may also work for tumours with U1 mutations. Alternatively, one might also target specific mis-spliced isoforms—such as the cell-surface protein CD44—via oligonucleotides or antibodies<sup>26,28</sup>. Genomic regions such as the U1 gene locus described here are generally overlooked in cancer sequencing studies. Future driver discovery studies that focus on these difficult regions might discover additional noncoding cancer drivers.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1651-z>.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

1. Khurana, E. et al. Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17**, 93–108 (2016).
2. Shuai, S., Gallinger, S. & Stein, L. D. DriverPower: combined burden and functional impact tests for cancer driver discovery. Preprint at <https://www.biorxiv.org/content/10.1101/215244v1> (2017).
3. Yoshida, K. et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64–69 (2011).
4. Wang, L. et al. *SF3B1* and other novel cancer genes in chronic lymphocytic leukemia. *N. Engl. J. Med.* **365**, 2497–2506 (2011).

5. Seiler, M. et al. Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. *Cell Reports* **23**, 282–296.e4 (2018).
6. Quesada, V. et al. Exome sequencing identifies recurrent mutations of the splicing factor *SF3B1* gene in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 47–52 (2011).
7. Denison, R. A., Van Arsdell, S. W., Bernstein, L. B. & Weiner, A. M. Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome. *Proc. Natl Acad. Sci. USA* **78**, 810–814 (1981).
8. Manser, T. & Gesteland, R. F. Human U1 loci: genes for human U1 RNA have dramatically similar genomic environments. *Cell* **29**, 257–264 (1982).
9. Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O. & Stein, L. D. Pan-cancer analysis of whole genomes. Preprint at <https://www.biorxiv.org/content/10.1101/162784v1> (2019).
10. Rheinbay, E. et al. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. Preprint at <https://www.biorxiv.org/content/10.1101/237313v1> (2017).
11. Kondo, Y., Oubridge, C., van Roon, A.-M. M. & Nagai, K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *eLife* **4**, e04986 (2015).
12. Suzuki, H. et al. Recurrent noncoding U1-snRNA mutations drive cryptic splicing in SHH medulloblastoma. *Nature* <https://doi.org/10.1038/s41586-019-1650-0> (2019).
13. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
14. Wang, L. et al. Transcriptomic characterization of *SF3B1* mutation reveals its pleiotropic effects in chronic lymphocytic leukemia. *Cancer Cell* **30**, 750–763 (2016).
15. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
16. Herishanu, Y., Gibellini, F., Njuguna, N., Keyvanfar, K. & Wiestner, A. CD44 signaling via PI3K/AKT and MAPK/ERK pathways protects CLL cells from spontaneous and drug induced apoptosis. *Blood* **112**, 541 (2008).
17. Fedorchenko, O. et al. CD44 regulates the apoptotic response and promotes disease development in chronic lymphocytic leukemia. *Blood* **121**, 4126–4136 (2013).
18. Popp, M. W. & Maquat, L. E. Leveraging rules of nonsense-mediated mRNA decay for genome engineering and personalized medicine. *Cell* **165**, 1319–1322 (2016).
19. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
20. Darman, R. B. et al. Cancer-associated *SF3B1* hotspot mutations induce cryptic 3' splice site selection through use of a different branch point. *Cell Reports* **13**, 1033–1045 (2015).
21. Kipps, T. J. et al. Chronic lymphocytic leukaemia. *Nat. Rev. Dis. Primers* **3**, 16096 (2017).
22. Nadeu, F. et al. Clinical impact of clonal and subclonal *TP53*, *SF3B1*, *BIRC3*, *NOTCH1*, and *ATM* mutations in chronic lymphocytic leukemia. *Blood* **127**, 2122–2130 (2016).
23. Llovet, J. M. et al. Hepatocellular carcinoma. *Nat. Rev. Dis. Primers* **2**, 16018 (2016).
24. Zhang, J. et al. Disease-associated mutation in *SRSF2* misregulates splicing by altering RNA-binding affinities. *Proc. Natl Acad. Sci. USA* **112**, E4726–E4734 (2015).
25. Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. K. RNA splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer* **16**, 413–430 (2016).
26. Lee, S. C.-W. & Abdel-Wahab, O. Therapeutic targeting of splicing in cancer. *Nat. Med.* **22**, 976–986 (2016).
27. Seiler, M. et al. H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. *Nat. Med.* **24**, 497–504 (2018).
28. Zhang, S. et al. Targeting chronic lymphocytic leukemia cells with a humanized monoclonal antibody specific for CD44. *Proc. Natl Acad. Sci. USA* **110**, 6127–6132 (2013).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019



## Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

### Data collection

All samples used in this study were from participants recruited and anonymized by individual International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) projects. Written informed consent was obtained from all human participants through individual projects. The PCAWG dataset consisting of 2,583 donors across 37 tumour types was collected from the ICGC Data Coordination Center (ICGC DCC). For whole-genome sequencing, all tumour and paired normal aligned BAMs ( $n = 5,166$ ) were retrieved. The use of PCAWG data was approved by the University of Toronto Research Ethics Board under RIS Human Protocol Number 30278 and protocol title 'Pan-cancer Analysis of Whole Genomes: PAWG'.

For CLL, a total of 141 normal–tumour paired whole-genome sequencing and 299 RNA-seq from 318 donors were used in this study (122 donors with both). This cohort included 29 cases of high-count monoclonal B cell lymphocytosis (MBL) of the CLL-type, an early phase of overt CLL. In addition to data from the PCAWG lymph-CLL cohort ( $n = 90$ ) that originated from the CLLE-ES (Chronic Lymphocytic Leukaemia – Spain) project, we also incorporated additional CLLE-ES data deposited in the European Genome-phenome Archive (EGA) and ICGC DCC (data release 27). All samples in this cohort had been studied before any treatment. The specific clinical and biological data of these cases have previously been described<sup>29</sup>. The use of genomic data, clinical data and CLL samples was approved by the Hospital Clinic of Barcelona Institutional Review Board under protocol number HCB/2015/0814 and protocol title 'Functional and Clinical Impact of Genomic Analysis in CLL'.

For HCC, 315 normal–tumour paired whole-genome sequencing and 387 tumour RNA-seq from 613 HCC donors were used in total (89 donors had both forms of data available). The PCAWG liver-HCC cohort ( $n = 315$ ) included data from four projects: LICA-FR ( $n = 5$ ; Liver Cancer - France), LINC-JP ( $n = 28$ ; Liver Cancer - National Cancer Center, Japan), LIRI-JP ( $n = 229$ ; Liver Cancer - RIKEN, Japan) and LIHC-US ( $n = 53$ ; Liver Hepatocellular Carcinoma - TCGA, US). Additional HCC samples from the LIHC-US project were collected from the National Cancer Institute Genomic Data Commons (NCI GDC).

All genomic data included in this study used GRCh37 as the reference genome and GENCODE v19 as the reference gene annotation<sup>30</sup>.

### Mutation calling for U1

First, samples without enough coverage were flagged as genotype-undetermined and left to manually review. The coverage was determined by the median read depth at the 5' splice-site recognition sequence of seven U1 genes. For 2,434 donors with enough coverage ( $\geq 15$  median coverage in at least five U1 genes), all reads mapped by BWA MEM to U1 genes and pseudogenes as well as their flanking 1-kb regions were extracted with samtools and saved as miniBAMs<sup>31</sup>. These miniBAMs were then converted into paired FASTQ files and re-aligned with Bowtie2 (v.2.3.4.1) to GRCh37 in multiple mapping report mode (-k)<sup>32</sup>. Non-default parameters for Bowtie2 were '-score-min L,-0.3,-0.3-no-mixed-no-discordant -k 100-very-sensitive'. Then, for each pair of multiple mapped reads, only alignments with minimal total edit distance (sum of edit distance in two mates) were kept. Reads mapped to U1 pseudogenes or other genomic regions were discarded. Next, for each re-aligned BAM, we counted the number of variant reads and the read depth (number of reference reads + number of variant reads) for each position, and for forward and reverse strand separately. To account for multiple mapping, we performed an extra procedure only for the read depth counting: that is, when a read had  $k$  equally good alignments, we only counted it as  $1/k$  read. We then used a beta-binomial error model trained on a project-specific panel

of normal samples to call mutations, which was implemented with a modified version of EBCall<sup>33</sup>. Finally, we used IGV to manually curate all mutation calls and filtered out mutations that were supported by reads with multiple mismatches in the same gene, or that had three or more variant reads in the paired normal sample according to BWA MEM or Bowtie2 alignments<sup>34</sup>. To further minimize the false-negative rate for the g.3A>C mutation, we also assigned tumours that were called as wild type but that had two or more variant reads at the third base of any U1 genes to the undetermined group.

### RNA-seq data processing

To analyse additional samples and PCAWG samples together, we uniformly processed additional CLL and HCC RNA-seq data with a slightly modified version of the PCAWG RNA-seq STAR 2-pass pipeline<sup>35,36</sup>. To maximize the sensitivity of novel junction discovery, we added a customized junction file that was extracted from PCAWG STAR alignments. Gene-level expression was counted by htseq-count (v.0.9.1)<sup>37</sup>. Transcript-level expression was estimated by Kallisto (v.0.44.0)<sup>38</sup>. The quality control process was done with FastQC (v.0.11.7) and multiQC (v.1.5)<sup>39</sup>. The transcript integrity number was calculated with RSeQC (v.2.6.4)<sup>40</sup>. In this study, we kept only RNA-seq data that met the following criteria: first, FASTQ files passed at least three main FastQC flags (overrepresented sequences, per base N content, per base sequence quality, per sequence GC content and per sequence quality scores); second, more than 50% reads were uniquely mapped and the total number of reads mapped by STAR was greater than 1 million; third, the total number of fragments counted by htseq-count was greater than 5 million; and fourth, the transcript integrity number was greater than 50.

### Differential splicing and expression analysis

For intron-centric differential splicing analysis, the LeafCutter package was used to quantify intron usage and identify differentially spliced intron clusters between two conditions<sup>13</sup>. Splice junction files (SJ.out.tab) generated by STAR were used as input for LeafCutter. Only splice junctions supported by uniquely mapped reads and with at least 6-bp maximum overhang were used. An intron was considered as significantly differentially spliced when  $q < 0.1$  and absolute  $\log_2$ (effective size)  $> 1$ . The limma package was used for differential expression analysis<sup>41</sup>. Gene-level expression from htseq-count was used as input. A gene was considered as significantly differentially expressed when  $q < 0.1$  and absolute  $\log_2$ -transformed fold change  $> 1$ . For CLL, we used the IGHV status (U-CLL or M-CLL) as the covariate, and compared 11 tumours with U1 mutation and wild-type *SF3B1* and 26 tumours with wild-type U1 and *SF3B1* mutation with 254 tumours with wild-type U1 and *SF3B1*. For HCC, we used project code (LIRI-JP or LIHC-US) as the covariate to control for batch effects and compared 20 tumours with U1 mutation with 367 tumours with wild-type U1. We also used randomized comparisons as controls by permuting 'MUT' and 'WT' labels for both differential splicing and differential expression analysis.

We also performed exon-centric differential splicing analysis for U-CLL (6 cases with U1 mutation and wild-type *SF3B1*, 6 cases with wild-type U1 and *SF3B1* mutation, and 30 cases with wild-type U1 and *SF3B1*) using the rMATS package (v.4.0.2) with default parameters<sup>42</sup>. Differential splicing events with  $q < 0.1$  and absolute  $\Delta\text{PSI} > 0.1$  were considered as significant.

### Inference of U1 g.3A>C status

Separate models were built for CLL and HCC. Cross-validations were used to compare different models and settings (Supplementary Note). For CLL, we used RNA-seq data for 7 cases with U1 mutation and 60 cases with wild-type U1 as training data, and RNA-seq data from 232 cases as test data. For HCC, we used RNA-seq for 10 cases with U1 mutation and 60 cases with wild-type U1 as training data, and RNA-seq data from 317 cases as test data. For splicing-based models, training data were used to identify differentially spliced introns (3,174 features for CLL and



# Article

600 features for HCC) and the use of these introns was then used to train a random forest classifier with 100 trees. For expression-based models, training data were used to identify differentially expressed genes (502 features for CLL) and normalized expression of these genes was then used to train a random forest classifier with 100 trees. Finally, *t*-distributed stochastic neighbour embedding (*t*-SNE) was used to verify and visualize all predictions<sup>43</sup>.

## Calculation of *ABCD3* splice score

The *ABCD3* splice score for CLL was built with the number of uniquely mapped RNA-seq reads that support cryptic splice junctions ( $n_{\text{cryptic}}$ ) or annotated splice junctions ( $n_{\text{annotated}}$ ). In total, one annotated junction (chr1: 94,946,163–94,948,725) and four cryptic junction (chr1: 94,946,163–94,948,144, chr1: 94,946,163–94,946,964, chr1: 94,948,575–94,948,725 and chr1: 94,947,112–94,948,144) were used. Then, the score was calculated as follows:

$$ABCD3 \text{ splice score} = -\log_{10}\left\{n_{\text{annotated}}/(n_{\text{annotated}} + n_{\text{cryptic}})\right\}$$

A high score ( $\geq 1$ ) indicated that the patient with CLL was a carrier of the *g.3A>C* mutation.

## PCR-based SNP assay

Genomic DNA from 298 primary samples was tested using custom rhAmp SNP assays (Integrated DNA Technology). In brief, locus and allele-specific primers were generated individually for *RNU1\_batch* (*RNU1-1*, *RNU1-2*, *RNU1-3*, *RNU1-4* and *RNU1-18*) and *RNU1\_pseudo* (*RNU1-27P* and *RNU1-28P*). Assays were run in technical triplicate in 5  $\mu$ l volume (DNA concentration sampled at least 10 ng), with control gBlocks for wild-type, mutant and heterozygous genotypes. Reporter mix used Yakima Yellow (mutant) and FAM (wild-type) dyes as well as ROX dye for passive reference. Plates were read on the StepOnePlus (Applied Biosystems) RT-PCR machine, and genotypes called using the StepOne v.2.3 software. The primer sequences are available in Supplementary Table 6.

## Cell lines and exogenous expression of the U1 *g.3A>C* mutation

The pLKO.1-puro U6 sgRNA BfuAI stuffer lentiviral vector (Addgene) was modified by removing the internal U6 promoter (between NdeI and EcoRI), and replacing it with the U1 locus, including 393 bases of internal promoter, the U1 sequence and 39 bases of 3'-flanking region using the following oligonucleotides (U1-For\_EcoRI: 5'-GTCGAGAATTCTTG-GCGTACAGTCTGTTTTTG; U1-Rev\_NdeI: 5'-CTATCATATGTAAGGAC-CAGCTTCTTTGGGA). The *g.3A>C* mutation was introduced by PCR with the following oligonucleotides (U1-A\_C-for: 5'-GCCAGTAAGGATGAGA-TCTTCGGG; U1-A\_C-rev: 5'-CCCGAAGATCTCATCCTTACCTGGC) in combination with the corresponding previous primers. The PCR products were digested with NdeI and EcoRI, and cloned in the modified pLKO.1 plasmid. All plasmids were verified by Sanger sequencing.

CLL cell lines JVM3 and HG3 were grown in RPMI 1640, 10% FBS, 1% PSG and 1% AA; MEC1 was grown in IMDM, 10% FBS, 1% PSG and 1% AA; and the HEK-293T cell line was grown in DMEM, 10% FBS, 1% PSG. CLL cell lines HG3, MEC1 and JVM3 were obtained from DSMZ (<https://www.dsmz.de/catalogues/catalogue-human-and-animal-cell-lines.html>). The authenticity of the cell lines was tested with the AmpFLSTR Identifier Plus PCR Amplification Kit. CLL cell lines have tested negative for mycoplasma. For the production of lentiviral particles, the protocol from the manufacturer (Addgene) was used with minor modifications. Thus, HEK-293T cells ( $5 \times 10^6$  cells) were cultured in 10-cm plates and transfected using Lipofectamine Plus (Invitrogen) with 2  $\mu$ g of either pLKO.1-U1<sup>wt</sup> (containing the wild-type U1 locus) or pLKO.1-U1<sup>g.3A>C</sup> (containing the *g.3A>C* mutation), together with 1  $\mu$ g of psPAX2 packaging plasmid and 1  $\mu$ g of pMD2.G envelope plasmid. Twelve hours after transfection, the medium was replaced with complete medium, and 24 h later 10 ml of supernatant were filtered (0.45  $\mu$ m), and 4 ml was used to infect

CLL cell lines in the presence of 8  $\mu$ g/ml polybrene. The infection was repeated 24 h later, and after 24 h cells were plated in complete medium for one day, and then selected with 1.2  $\mu$ g/ml of puromycin. Cells were selected for four days, and total RNA was extracted with the Trizol method.

## Verification of the expression of the U1 *g.3A>C* mutation by 5' rapid amplification of cDNA ends

Rapid amplification of cDNA ends (RACE) was performed using 1  $\mu$ g of total RNA from JVM3, HG3 or MEC1 cell lines infected with either pLKO.1-U1<sup>wt</sup> or pLKO.1-U1<sup>g.3A>C</sup> following the recommendations of the manufacturer (Sigma-Aldrich), and the following specific oligonucleotides (U1-RACE\_SP1: 5'-CAGGGGAAAGCGCGAACGCAGT; U1-RACE\_SP2: 5'-CCCACTACCACAAATTATGC). A single amplification band of the expected size (160 bp) was excised from the gel, purified and sequenced with the internal oligonucleotide U1-RACE\_SP2.

## RNA-seq and data analysis for CLL cell lines

In total, 12 libraries—including 2 technical replicates for each of the 3 cell lines (JVM3, HG3 or MEC1) and 2 conditions (mutation or wild type)—were prepared as stranded total RNA-seq libraries and then sequenced with the Illumina HiSeq 4000 system ( $2 \times 76$  bp) with >40 million paired-end reads per sample. Cell line RNA-seq data were processed and analysed the same way as were primary tumour RNA-seq data. For differential splicing analysis, the same set of intron clusters used in primary CLL was tested in cell lines, so that their results were directly comparable. For the overlap test, a one-tailed Fisher's exact test was used.

## RT-PCR and qPCR validation of mis-splicing events in primary CLL

For PCR with reverse transcription (RT-PCR), RNA was obtained for samples from 14 patients with CLL, including 6 cases of U-CLL with U1 mutation, 4 cases of U-CLL with wild-type U1 and 4 cases of M-CLL with wild-type U1. cDNA was synthesized using the iScript cDNA Synthesis Kit (Bio-Rad 1708890). PCRs were performed using 1  $\mu$ l cDNA and the Taq PCR Master Mix Kit (Qiagen 201445) using 35 cycles. Products were run on the QIAxcel Advanced System (Qiagen). For quantitative PCR (qPCR), the same set of CLL samples was used except that one case of U-CLL with U1 mutation was exhausted. qPCR was performed using 1  $\mu$ l cDNA and the PowerUp SYBR Green Master Mix (Applied Biosystems A25742) in duplicates in a StepOnePlus Real-Time System (Applied Biosystems). Relative quantification was analysed with the  $2^{-\Delta\Delta C_t}$  method using *GAPDH* as the endogenous control. The primer sequences are available in Supplementary Table 6.

## Gene-set overrepresentation and enrichment analysis

We identified gene lists that were significantly overrepresented in differentially spliced genes from Gene Ontology, Kyoto Encyclopedia of Genes and Genomes and Reactome databases using g:Profiler<sup>44–47</sup>. We also conducted gene-set enrichment analysis (GSEA) for differentially expressed genes using pre-ranked gene lists ordered by  $-\log_{10}(P \text{ value}) \times (\text{sign of fold change})$ <sup>19</sup>. Both classical and weighted enrichment statistics were used in GSEA. For GSEA, we focused on C2 (curated) and C5 (Gene Ontology) gene sets in the Molecular Signatures Database (MSigDB v.6.2)<sup>48</sup>.

## Mutual-exclusivity and co-occurrence analysis

We collected lists of CLL and HCC driver alterations from the literature<sup>23,29</sup>. All CLL samples and whole-genome-sequenced HCC samples were used in the analysis. HCC samples were analysed separately based on project (LIRI-JP or LIHC-US), and as a combined cohort. To determine the pairwise significance between the U1 mutation and other driver events, we used the Cochran–Mantel–Haenszel  $\chi^2$  test for the combined HCC cohort, and Fisher's exact test for each project. As the detection of *TERT* promoter mutations was underpowered in many PCAWG HCC

samples (especially for the LIRI-JP project), we also included rescued *TERT* promoter mutations as previously described<sup>49</sup>.

### Clinical data analysis

All clinical data analysed here have previously been described<sup>29,50,51</sup>. Patient outcomes were analysed with the log-rank test for a single variate and Cox proportional hazards regression model for multivariate. For CLL, we analysed overall survival and time to first treatment from the time of sampling. Cases with MBL were not included in the outcome analysis. For HCC in LIRI-JP, we only analysed overall survival. For HCC in LIHC-US, we analysed two endpoints (overall survival and progression-free interval) as recommended by the TCGA PanCancer Atlas<sup>52</sup>. Age at diagnosis between mutated and unmutated groups was tested with two-sample Wilcoxon rank-sum tests. The association between U1 mutations and categorical patient characteristics (such as gender, IGHV status, infection with hepatitis B or C virus and alcohol history) were analysed with Fisher's exact test. For alcohol history, two HCC projects used different indicators. For LIHC-US, we used binary alcoholic liver disease history. For LIRI-JP, we collapsed its four-level alcohol intake indicators (a, no alcohol intake; b, social drinker; c, about 60 g every day; d, 60 g and more every day) into a binary factor (0, a and b; 1, c and d).

### Statistical analysis

All statistical tests were two-sided unless otherwise stated. All statistical methods are described in the corresponding sections and  $P < 0.05$  was considered as significant when only a single test was performed. All false-discovery rate controls were conducted with the Benjamini–Hochberg procedure and false-discovery rate of 10% ( $q < 0.1$ ) was selected as the significant threshold.

### Code availability

All published computational programs used in this study are indicated in corresponding sections. Scripts used to perform U1 snRNA mutational calling are available at <https://github.com/smsuail/U1-snRNA>.

### Data availability

PCAWG data are available at ICGC DCC (<https://docs.icgc.org/pcawg/data/>; donor identifiers in Supplementary Table 1). Additional CLL data (donor identifiers in Supplementary Table 3) are available at ICGC DCC ([https://dcc.icgc.org/releases/release\\_27/Projects/CLLE-ES](https://dcc.icgc.org/releases/release_27/Projects/CLLE-ES)) and EGA (raw data under accession numbers EGAS00001000374 and EGAS00001001306). Additional HCC data are available at GDC Data Portal (raw and processed data under project code TCGA-LIHC; donor identifiers in Supplementary Table 3). CLL cell line RNA-seq data are available at GSE134197.

29. Puente, X. S. et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
30. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
31. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
32. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
33. Shiraishi, Y. et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41**, e89 (2013).
34. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the Integrative Genomics Viewer. *Cancer Res.* **77**, e31–e34 (2017).
35. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

36. PCAWG Transcriptome Core Group et al. Genomic basis for RNA alterations revealed by whole-genome analyses of 27 cancer types. Preprint at <https://www.biorxiv.org/content/10.1101/183889v2> (2018).
37. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
38. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
39. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
40. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
41. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
42. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-seq data. *Proc. Natl Acad. Sci. USA* **111**, E5593–E5601 (2014).
43. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
44. Reimand, J. et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–W89 (2016).
45. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
46. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
47. Croft, D. et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
48. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
49. Zhang, Y. et al. Whole genome and RNA sequencing of 1,220 cancers reveals hundreds of genes deregulated by rearrangement of cis-regulatory elements. Preprint at <https://www.biorxiv.org/content/10.1101/099861v3> (2017).
50. Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
51. The Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327–1341.e23 (2017).
52. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416.e11 (2018).

**Acknowledgements** The authors acknowledge the use of pre-embargo whole-genome sequencing alignment data from the PCAWG project, approved by the PCAWG Steering Committee (with L.D.S. recused). This work was supported by the Government of Ontario (S.S. and L.D.S.), the Instituto de Salud Carlos III (project PMP15/00007; to F.N., M.P., J.D., X.S.P., C.L.-O. and E.C.), the 'la Caixa' Foundation Grant No HR17-00221 (Health Research 2017 Program; to F.N., M.P., J.D., X.S.P., C.L.-O. and E.C.) and the Ministerio de Economía y Competitividad (MINECO) SAF2013-45836-R (to X.S.P., A.D.-N. and A.G.-F.). A.D.-N. is supported by the Department of Education of the Basque Government (grant number PRE\_2017\_1\_0100). E.C. is supported by ICREA under the ICREA Academia programme. F.N. is supported by a pre-doctoral fellowship of the Ministerio de Economía y Competitividad (MINECO), BES-2016-076372). H.S. is a recipient of a Research Fellowship (Astellas Foundation for Research on Metabolic Disorders).

**Author contributions** S.S. and L.D.S. designed the experiments, interpreted results and prepared the manuscript with inputs from all authors. S.S. and H.S. performed primary tumour whole-genome sequencing and RNA-seq analysis (Figs. 1, 2, Extended Data Figs. 1, 2). S.A.K. and F.N. conducted rhAmp (Extended Data Fig. 1d) and CLL RT-qPCR experiments (Extended Data Fig. 4). S.S. and L.D.S. performed pathway and gene-set analysis (Fig. 3, Extended Data Fig. 3, 5), comparison with *SF3B1* (Extended Data Fig. 8), and clinical and driver analysis in the HCC cohort (Fig. 4b–d, f, g, Extended Data Figs. 9, 10d–g). F.N. and J.D. performed clinical and driver analysis in the CLL cohort (Fig. 4a, e, Extended Data Figs. 7, 10a–c). X.S.P., A.G.-F. and A.D.-N. conducted cell line experiments (Extended Data Fig. 6; data analysis of Extended Data Fig. 6b–h was performed by S.S.). E.C. and C.L.-O. assembled the cohorts and co-directed earlier studies that produced the CLL genomic and transcriptomic data used in Figs. 1–4, Extended Data Fig. 2–8. E.C., M.P., J.D. and C.L.-O. provided CLL tissue samples and the corresponding donor clinical data used in Fig. 4a, e, Extended Data Figs. 7, 10a–c. L.D.S., E.C. and M.D.T. supervised the project. All authors read, had the opportunity to comment on and have approved the manuscript.

**Competing interests** The authors declare no competing interests.

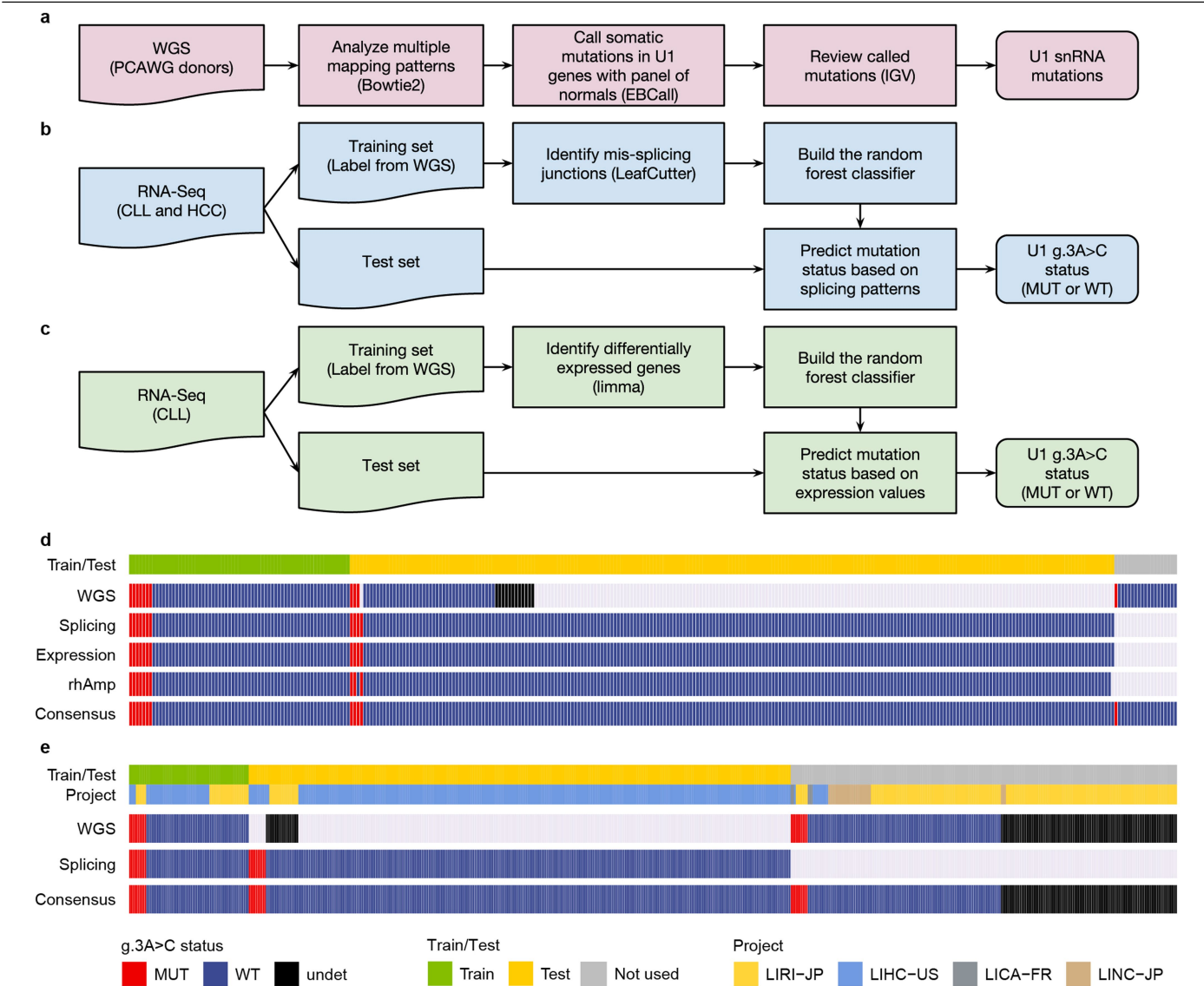
### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1651-z>.

**Correspondence and requests for materials** should be addressed to L.D.S.

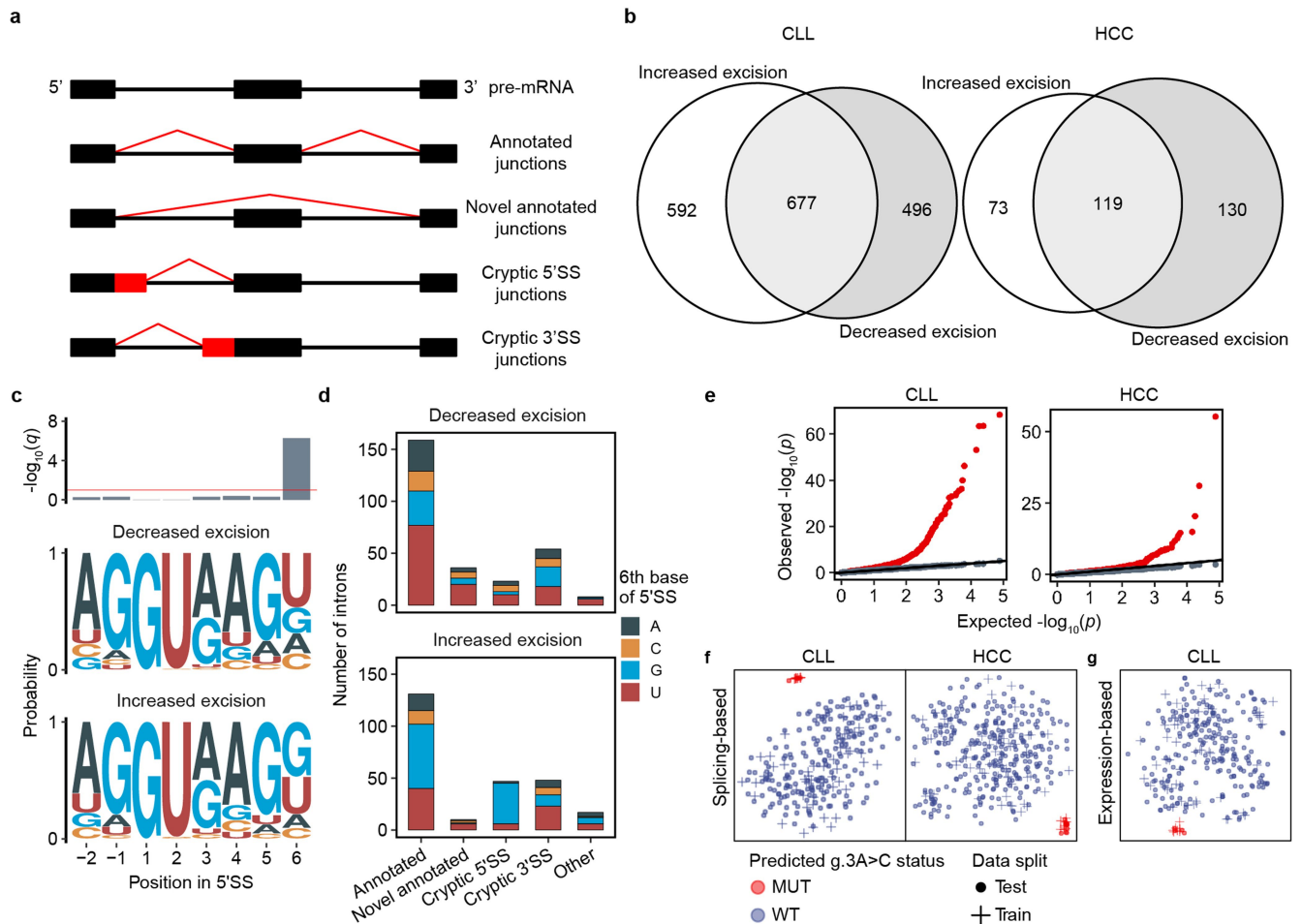
**Peer review information** *Nature* thanks Rotem Karni, Brandon Wainwright and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



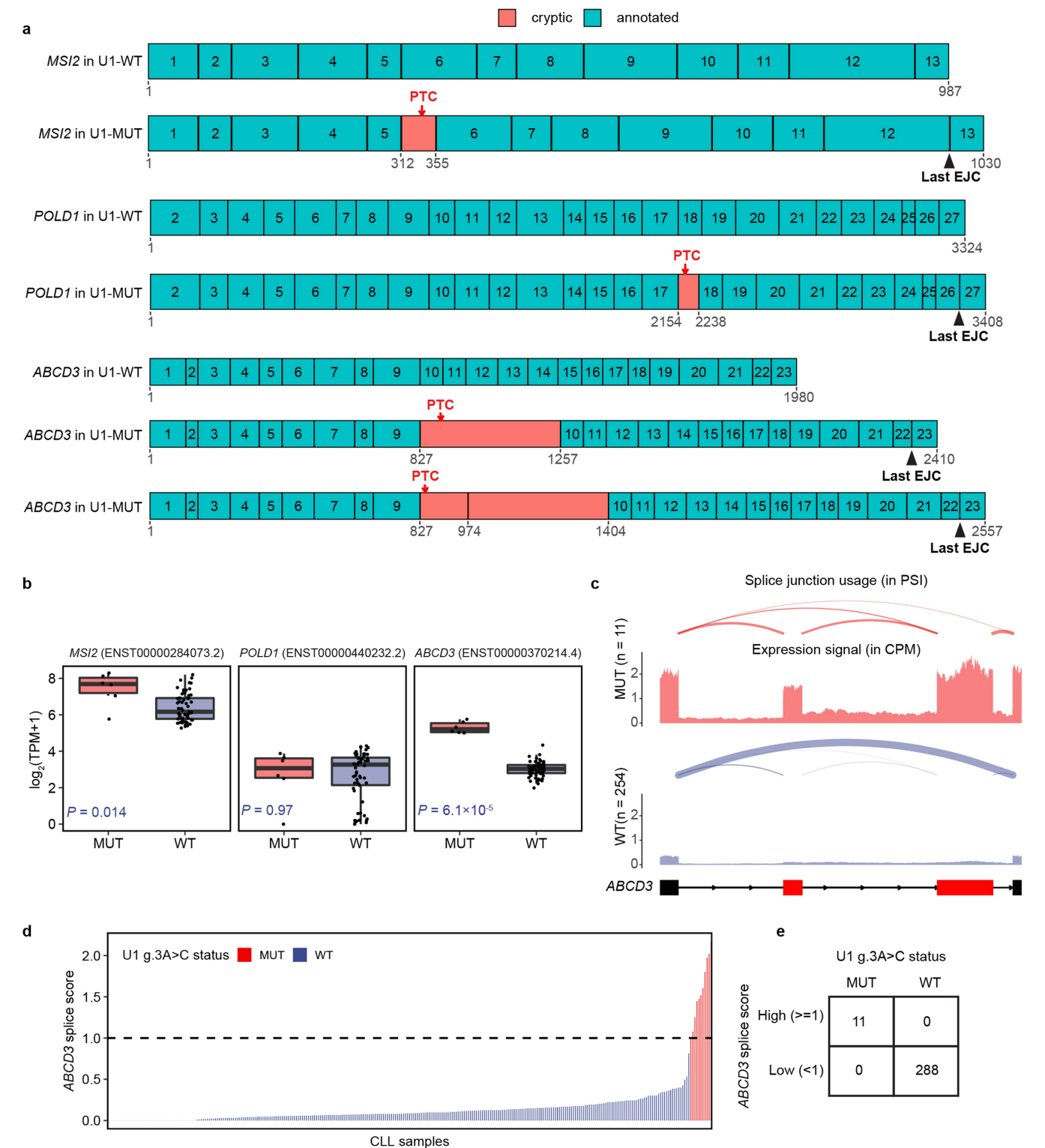
**Extended Data Fig. 1 | Overview of genotyping methods and results. a,** Computational pipeline used to identify somatic mutations in U1 genes from whole-genome sequencing data. **b,** Computational pipeline used to infer the U1 g.3A>C status on the basis of splicing patterns. **c,** Computational pipeline used

to infer the U1 g.3A>C status on the basis of expression patterns. **d, e,** U1 g.3A>C status for 318 CLL (**d**) and 613 HCC samples (**e**). Consensus genotypes are used in downstream analysis. One hundred and three undetermined HCC samples are excluded from downstream analysis. Grey indicates data not available.



**Extended Data Fig. 2 | Global transcriptomic changes in CLL and HCC. a,** Category of mis-spliced introns. Black boxes indicate exons and lines indicate introns; red lines indicate actual splice junctions. A junction is described as 'annotated' if the junction matches any annotated introns. A junction is described as 'novel annotated' if both splice sites are annotated but not paired. A cryptic 5' splice site junction exists if only the 3' splice site is annotated. A cryptic 3' splice site junction exists if only the 5' splice site is annotated. **b,** Euler plots of mis-spliced intron clusters in CLL and HCC. Increased excision and decreased excision represent intron clusters that have significantly mis-spliced introns with  $\Delta\text{PSI} > 0$  and  $\Delta\text{PSI} < 0$ , respectively. **c,** 5' splice site for introns with increased ( $n = 239$ ) or decreased excision ( $n = 294$ ) in HCC with U1 mutation. Top, bar plot

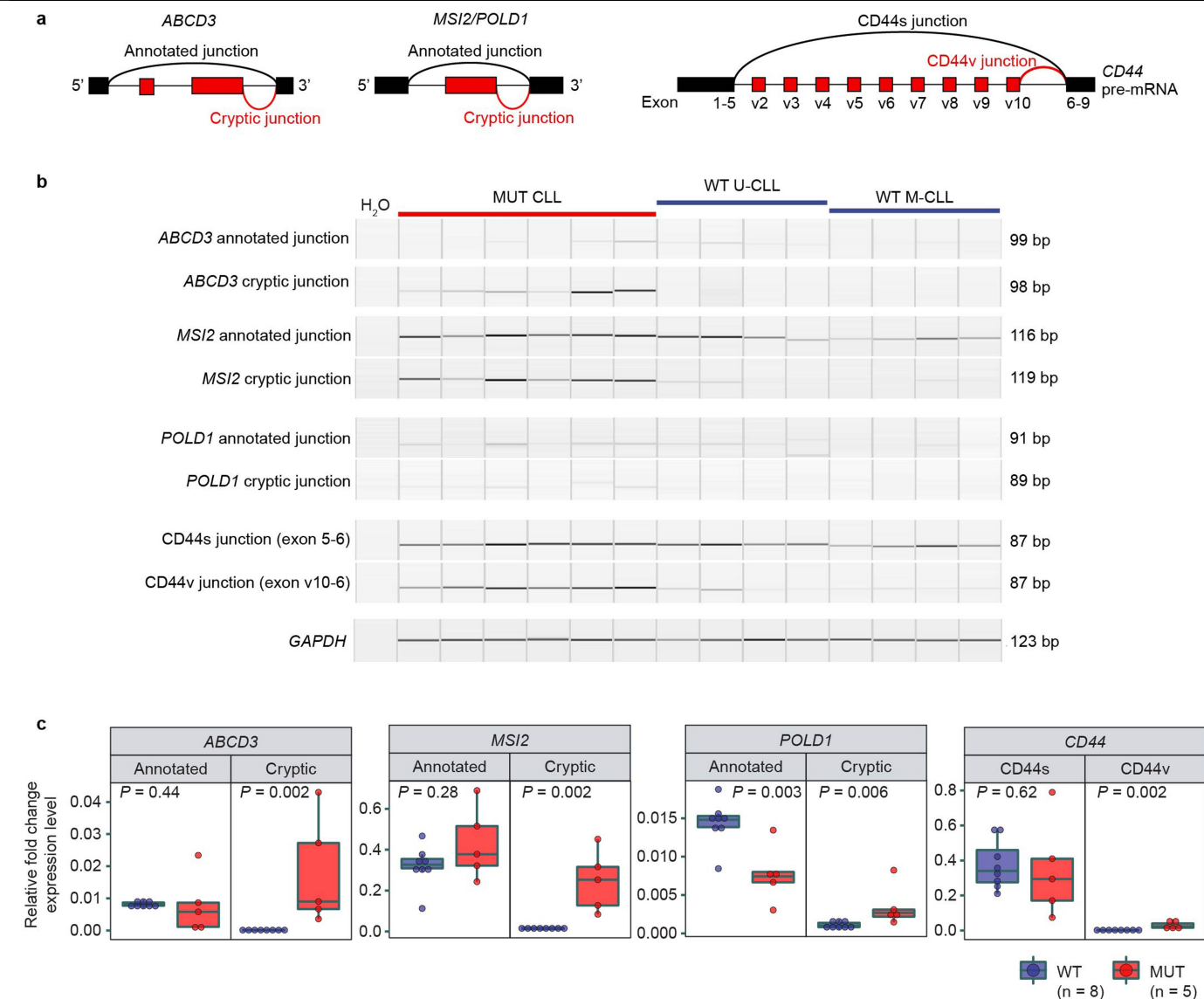
shows Benjamini-Hochberg adjusted  $P$  values from  $\chi^2$  tests; red line indicates  $q = 0.1$ . **d,** Category of mis-splicing events in HCC with U1 mutation. **e,**  $P$  value quantile-quantile plots for differential expression analysis. Global gene-expression changes cannot be detected in the randomized dataset.  $P$  values are from limma. Biologically independent patient samples are used for CLL (11 cases with U1 mutation versus 254 cases with wild-type U1) and HCC (20 cases with U1 mutation versus 367 cases with wild-type U1). **f, g,**  $t$ -SNE coordinates are computed from differentially spliced introns, and predictions are made with the splicing-based classifier. **g,**  $t$ -SNE coordinates are computed from differentially expressed genes, and predictions are made with the expression-based classifier.



**Extended Data Fig. 3 | Premature termination codon-bearing mis-spliced isoforms for *MSI2*, *POLD1* and *ABCD3* in CLL with U1 mutation.** **a**, Coding sequence of *MSI2* (ENST00000284073.2), *POLD1* (ENST00000440232.2) and *ABCD3* (ENST00000370214.4) in CLL. Annotated exons are numbered. Positions 1 to 3 indicate the start codon. The first stop codon in each cryptic exon is labelled as the premature termination codon (PTC). The last exon junction complex (EJC) is also labelled. The distance between premature termination codon and the last exon junction complex is >55 nt in all isoforms with cryptic exons. **b**, Expression box plots for three of the transcripts that bear premature termination codon, shown in **a**.  $P$  values are from two-sided Wilcoxon rank-sum

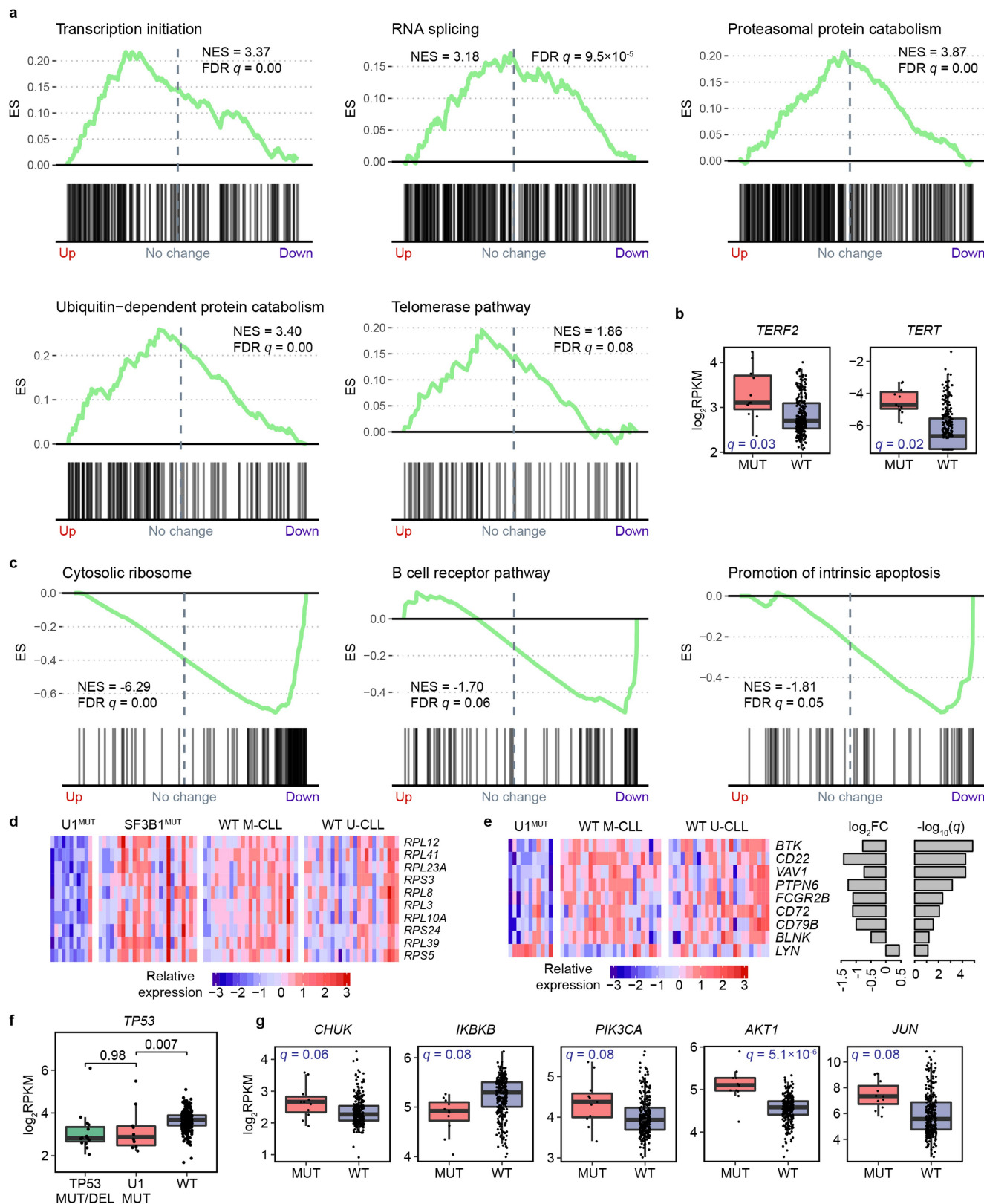
tests. MUT ( $n = 6$ ) and WT ( $n = 61$ ) represent independent CLL samples with or without the g.3A>C mutation, respectively. In the box plot, centre line, box limits, whiskers and points represent the median, 25th and 75th percentiles, 1.5 $\times$  interquartile range and individual samples, respectively. **c**, Sashimi plot for mis-splicing of *ABCD3*. This sashimi plot is the same as those in Fig. 3. **d**, Splice score of *ABCD3* across all CLL samples. The *ABCD3* splice score is calculated from RNA-seq data (Methods). **e**, Using the *ABCD3* splice score as a marker for U1 mutation in CLL. The *ABCD3* splice score can be used to predict the U1 g.3A>C status with 100% accuracy in CLL.





**Extended Data Fig. 4 | Validation of mis-splicing in *MSI2*, *POLD1*, *CD44* and *ABCD3*.** **a**, Junctions used in RT-qPCR validations. For *MSI2*, *POLD1* and *ABCD3*, one annotated junction and one cryptic junction are used. For *CD44*, one *CD44s* junction and one *CD44v* junction are used. Cryptic and *CD44v* junctions should be over-excised in samples with U1 mutation. **b**, RT-PCR result for junctions shown in **a**. MUT ( $n = 6$ ) and WT ( $n = 8$ ) represent independent CLL samples with and without the g.3A>C mutation, respectively. *GAPDH* is used as control.

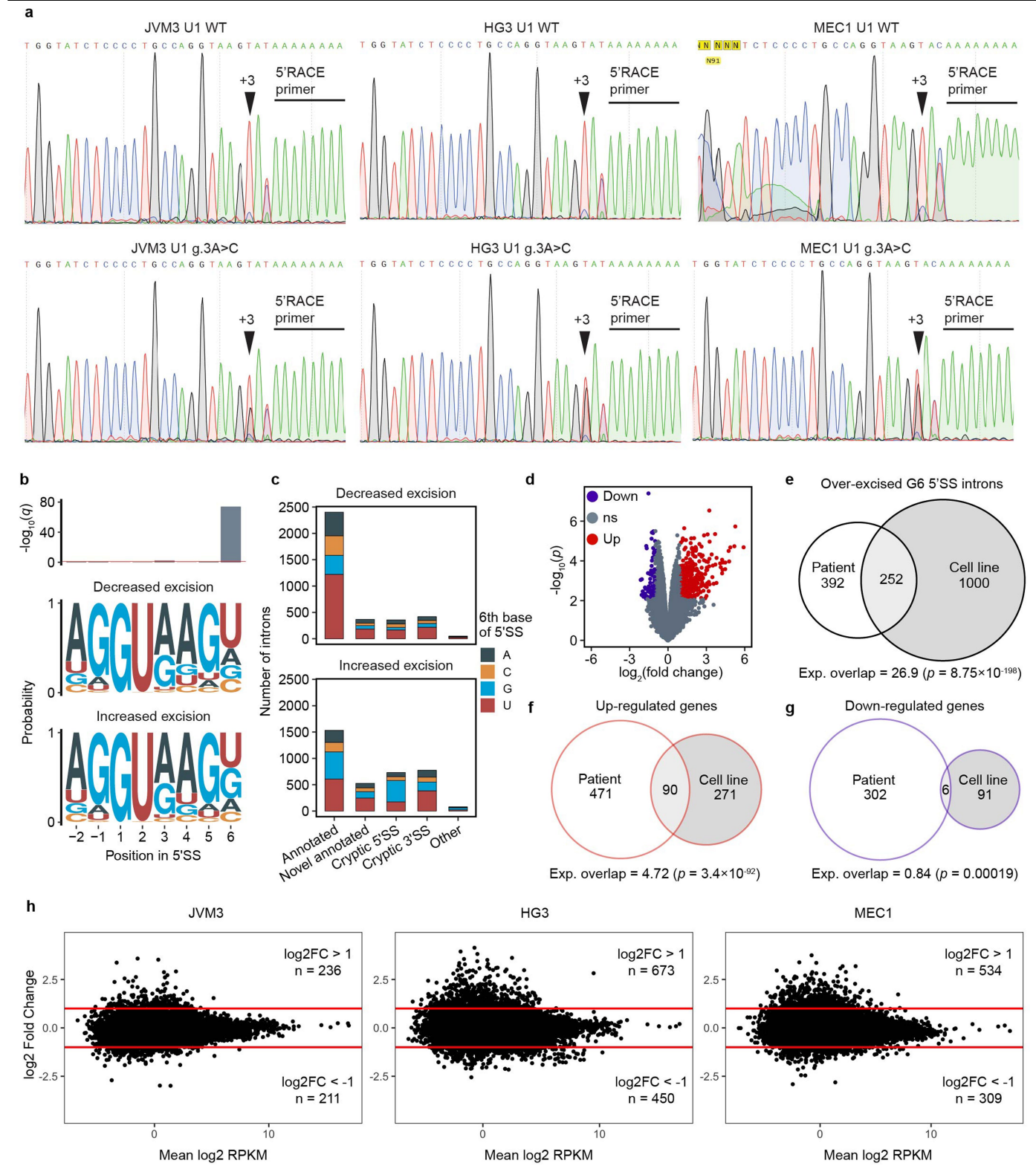
The length of all PCR products is shown on the right. The experiment was conducted once. For gel source data, see Supplementary Fig. 1. **c**, RT-qPCR result for junctions shown in **a**. Expression of cryptic and *CD44v* junctions are significantly higher in CLL samples with U1 mutation, as expected.  $P$  values are from two-sided Wilcoxon rank-sum tests. MUT ( $n = 5$ ) and WT ( $n = 8$ ) represent biologically independent CLL samples with and without the g.3A>C mutation, respectively.



Extended Data Fig. 5 | See next page for caption.

**Extended Data Fig. 5 | Genes and pathways altered in CLL with U1 mutation.**  
**a, c,** Representative upregulated (**a**) and downregulated (**c**) gene sets in CLL with U1 mutation ( $n = 11$  independent patients). For **a, c**, enrichment score (ES), normalized enrichment score (NES) and false-discovery rate (FDR)  $q$  are from the permutation test of GSEA. Genes in the  $x$  axis are sorted from the most significantly upregulated gene to the most significantly downregulated gene. The dashed line indicates fold change = 1. **b,** Expression box plots for *TERF2* and *TERT* in the telomere maintenance pathway.  $q$  values are from limma. MUT ( $n = 11$ ) and WT ( $n = 254$ ) represent biologically independent CLL samples with and without the g.3A>C mutation, respectively. **d, e,** Expression heat map of ribosomal genes and BCR genes. **d,** Only the top 10 differentially expressed ribosomal genes are shown. **e,** Right, bar charts show the  $\log_2$ -transformed fold change and  $-\log_{10}(q \text{ value})$  from limma ( $n = 11$  independent samples of CLL with

U1 mutation;  $n = 254$  independent samples of CLL with wild-type U1). **f,** Expression box plot for *TP53*. *TP53* MUT/DEL ( $n = 15$  independent patients) represents CLL samples with *TP53* mutation and/or deletion; U1 MUT ( $n = 10$  independent patients) represents samples with the g.3A>C mutation; WT ( $n = 273$  independent patients) represents the rest of the CLL samples. One sample with both *TP53* and g.3A>C mutation is excluded.  $P$  values are from two-sided Wilcoxon rank-sum tests. **g,** Expression box plots for genes in the NF- $\kappa$ B (*CHUK* and *IKBKB*), PI3K/AKT (*PIK3CA* and *AKT1*) and MAPK/ERK (*JUN*) pathways. These pathways are downstream effectors of the BCR pathways.  $q$  values are from limma. MUT ( $n = 11$ ) and WT ( $n = 254$ ) are the same as in **b**. In all box plots, centre line, box limits, whiskers and points represent median, 25th and 75th percentiles, 1.5 $\times$  interquartile range and individual samples, respectively.

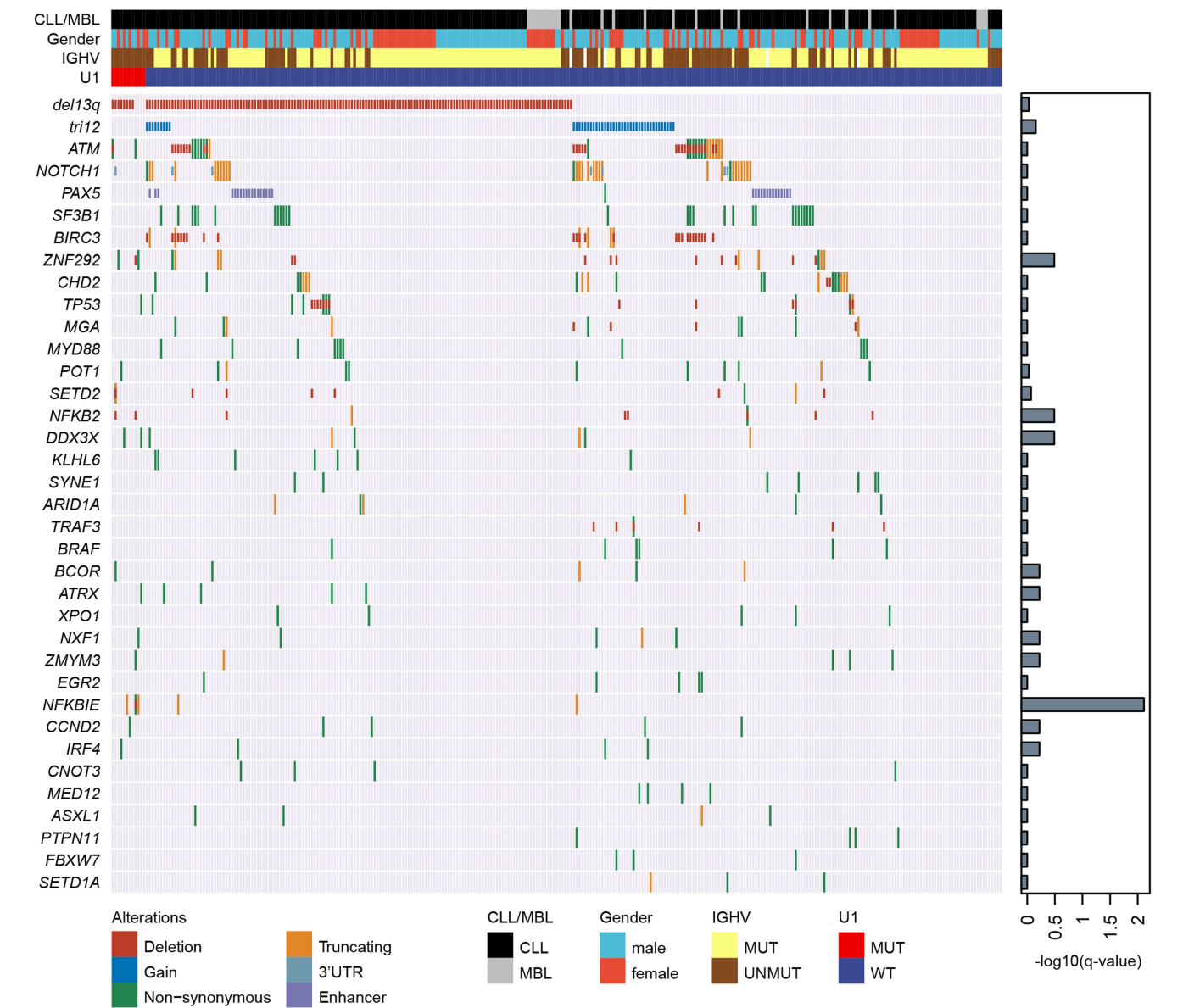


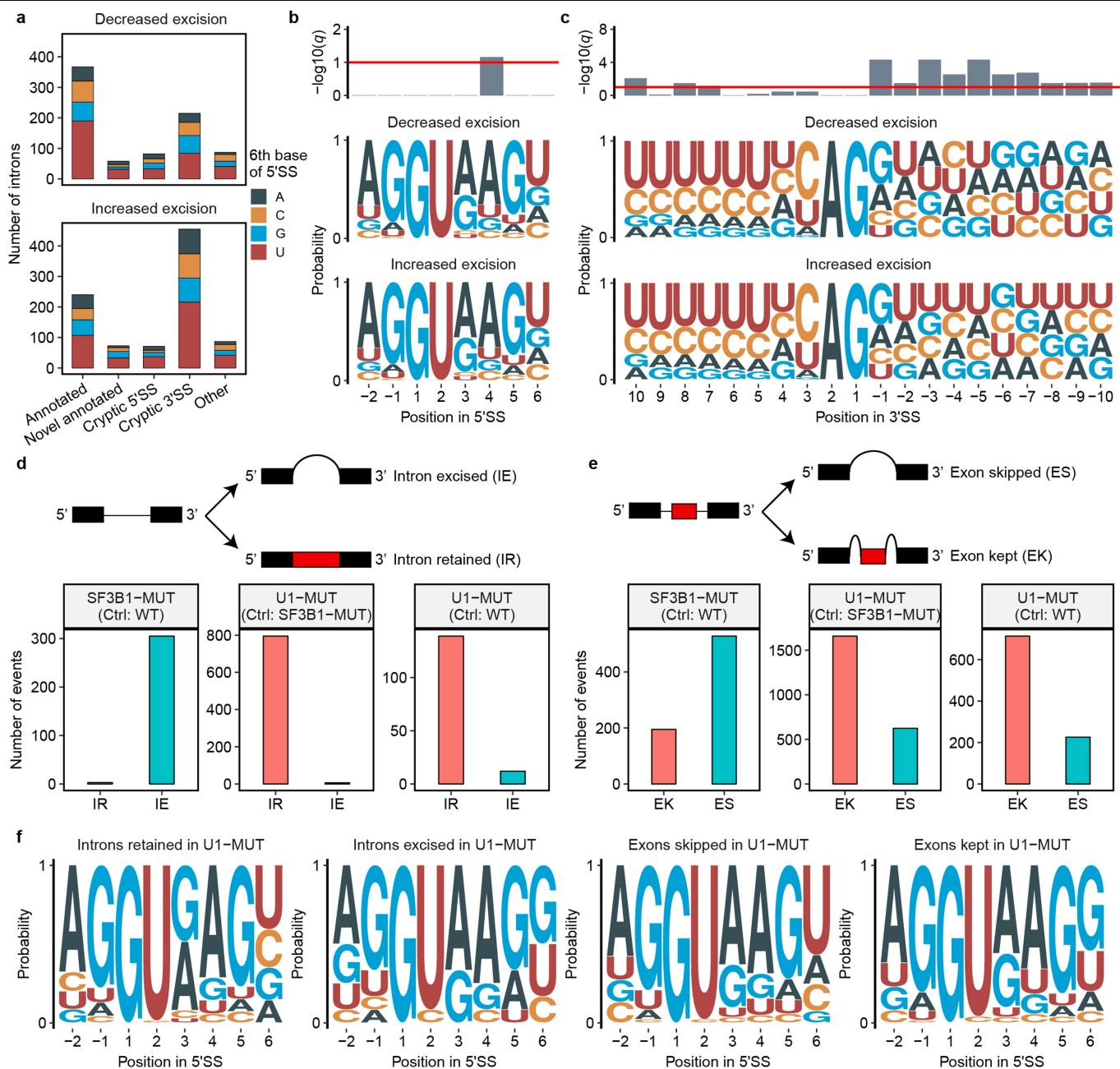
**Extended Data Fig. 6** | See next page for caption.

**Extended Data Fig. 6 | Exogenous expression of the g.3A>C mutation in cell lines induced transcriptome-wide changes.** **a**, 5' RACE confirming the expression of the U1 g.3A>C mutation in three CLL cell lines. CLL cell lines (JVM3, HG3 and MEC1) are infected with lentiviral particles that contain U1 locus with or without the g.3A>C mutation (g.3A>C or WT, respectively). The electropherograms correspond to the sequence of the PCR product (reverse-strand). The arrowheads indicate the location of the 3rd base of U1. The location of the 5' RACE primer is also indicated. The experiment was conducted once. **b**, 5' splice site for introns with increased or decreased excision in CLL cell lines with U1 mutation ( $n = 3$  biological independent cell lines). Top, bar chart shows  $q$  values from  $\chi^2$  tests for base composition difference; red line indicates the  $q = 0.1$  cutoff. **c**, Category of mis-splicing events in CLL cell lines. Extended Data

Figure 2a provides the definition of each category. The number of introns is coloured by the sixth base of 5' splice site. **d**, Volcano plot for differential expression analysis of CLL cell lines.  $q$  values are from limma.  $n = 3$  biological independent cell lines; each cell line has a form with (MUT) and without (WT) U1 mutation. **e–g**, Euler plots comparing differentially spliced introns (**e**) and differentially expressed genes (**f**, **g**) between primary CLL from patients and cell lines. **e**, Over-excised introns with the G6 5' splice site that are direct targets of the U1 mutation are compared.  $P$  values are from one-tailed Fisher's exact test. The expected number of overlaps is also shown. **h**, Asymmetric changes in gene expression for cell lines. In all three cell lines, more genes are upregulated ( $\log_2$ -transformed fold change  $> 1$ ) than downregulated ( $\log_2$ -transformed fold change  $< -1$ ).

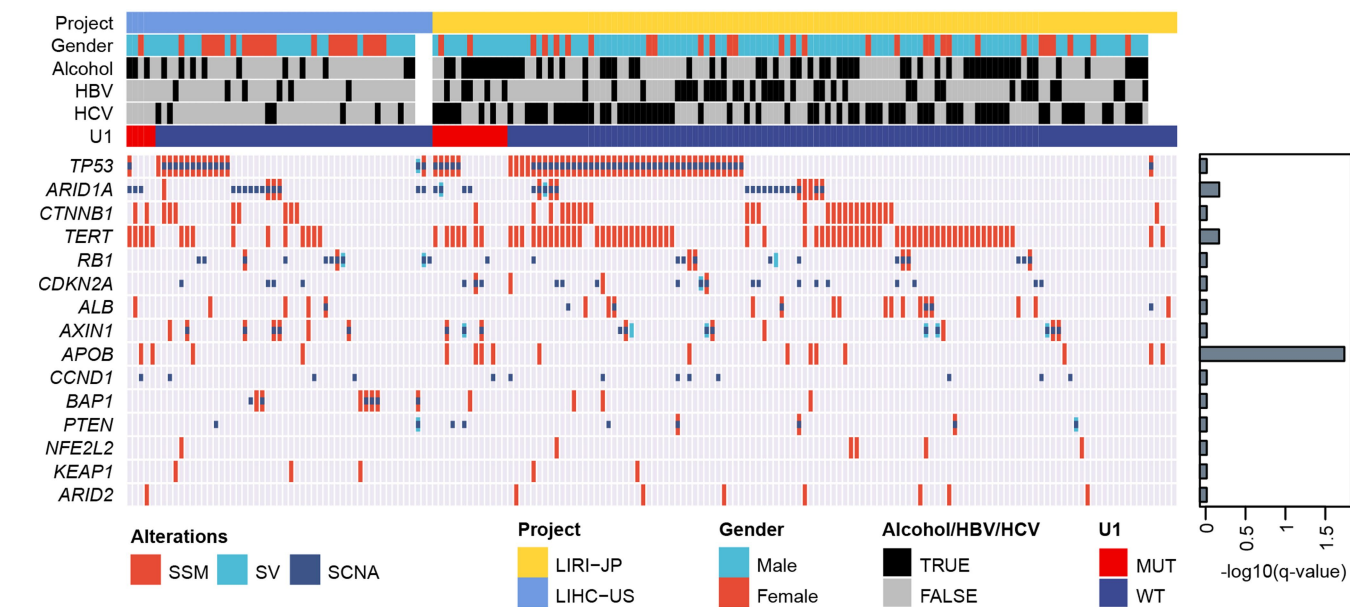






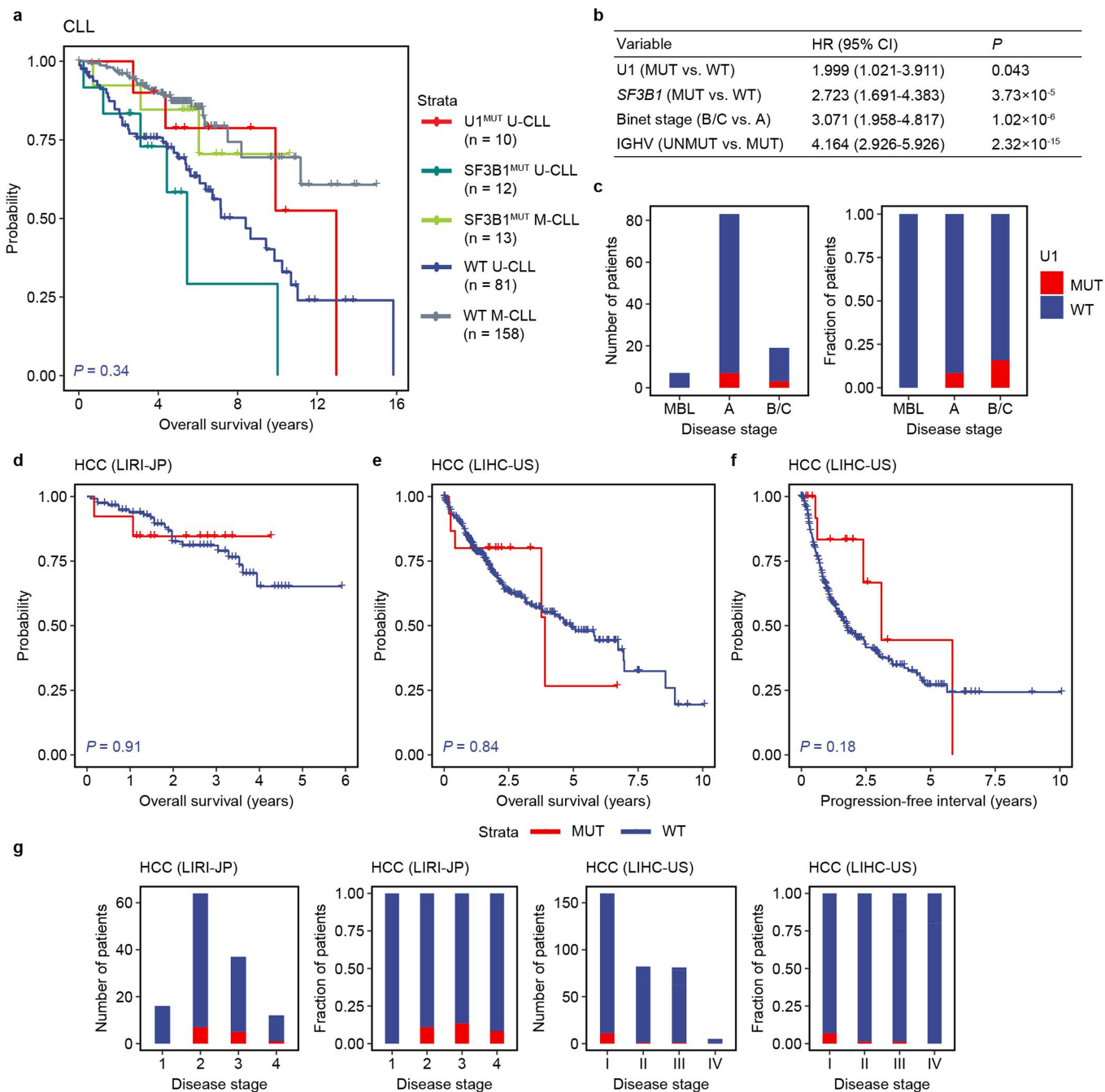
**Extended Data Fig. 8 | Consequence of *SF3B1* mutation and its relation to the U1 g.3A>C mutation in CLL.** **a**, Category of mis-splicing events in CLL with *SF3B1* mutation. Unlike the g.3A>C mutation, *SF3B1* mutation induces more cryptic 3' splice site changes. **b, c**, Sequence motifs of 5' splice site (**b**) and 3' splice site (**c**) for mis-spliced introns in CLL with *SF3B1* mutation ( $n = 6$ ). In **b, c**, the bar plots on the top show Benjamini-Hochberg adjusted  $P$  values from  $\chi^2$  tests. Red lines indicate the  $q = 0.1$  cutoff. **d, e**, Intron retention (**d**) and exon skipping (**e**) events

in CLL with U1 mutation or *SF3B1* mutation. In **d, e**, pairwise comparisons are made between three groups of U-CLL: with U1 g.3A>C mutation (U1-MUT,  $n = 6$ ); with *SF3B1* mutation (SF3B1-MUT,  $n = 6$ ); without U1 or *SF3B1* mutation (WT;  $n = 30$ ). The number of events is counted with respect to corresponding controls (ctrl). **f**, 5' splice site motifs for intron retention and exon skipping events in samples with U1 mutation, as compared to wild-type samples. The sequence motif for exon skipping events is from 5' splice site of the red exon in **e**.



**Extended Data Fig. 9 | Major driver events and clinical features of HCC.** Left, this oncoprint shows 15 major driver alterations across all whole-genome-sequenced samples from two HCC projects: LIHC-US ( $n = 53$ ) and LIRI-JP ( $n = 129$ ). Project code, gender, risk factor status (heavy alcohol use, or infection with hepatitis B or C virus) and U1g.3A>C status are shown on the top. White colour in

top annotations indicates that data are not available. Right, bar plot shows  $q$  values from Cochran–Mantel–Haenszel  $\chi^2$  tests that compare each alteration with the U1 mutation. Three types of alterations are used in the plot: simple somatic mutations (SSM), structural variations (SV) and somatic copy-number alterations (SCNA).



**Extended Data Fig. 10 | Clinical analysis of CLL and HCC.** **a**, Kaplan–Meier plot of overall survival for CLL.  $P$  value is from the Cox model. **b**, Independent prognostic value of the U1g.3A>C mutation in CLL. Multivariate Cox model results for time to first treatment in 301 patients with CLL (number of events = 153). HR, hazard ratio; CI, confidence interval. **c**, U1 mutation frequency by disease stage in CLL. A and B/C, Binet A and B/C stages. **d**, Kaplan–Meier plot of overall survival for HCC samples in the LIRI-JP project. MUT indicates samples with the g.3A>C mutation and WT indicates samples without the g.3A>C

mutation; 13 MUT and 116 WT samples are used. **e**, **f**, Kaplan–Meier plots of OS (**e**) and progression-free interval (**f**) for HCC samples in the LIHC-US project. MUT indicates samples with the g.3A>C mutation and WT indicates samples without the g.3A>C mutation; 15 MUT and 335 WT samples are used. For **d–f**,  $P$  values are from two-sided log-rank tests; **g**, U1 mutation frequency by disease stage in HCC. For LIRI-JP, staging uses the Liver Cancer Study Group of Japan system. For LIHC-US, staging uses the American Joint Committee on Cancer system.

## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](#)

Data collection

No code was used for data collection.

Data analysis

The following published softwares were used in this study: Bowtie2 (v2.3.4.1); STAR (v2.5.4); htseq-count (v0.9.1); Kallisto (v0.44.0); multiQC(v1.5); RSeQC (v2.6.4); FastQC (v0.11.7); rMATS (v4.0.2); leafCutter (v0.2.7); StepOne (v2.3); limma (v3.38.3). Customized code for mutation calling have been deposited into GitHub (<https://github.com/smsuui/U1-snRNA>).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

PCAWG data are available at ICGC DCC (<https://docs.icgc.org/pcawg/data/>). Additional CLL data are available at ICGC DCC ([https://dcc.icgc.org/releases/release\\_27/Projects/CLLE-ES](https://dcc.icgc.org/releases/release_27/Projects/CLLE-ES)) and EGA (raw data under accession numbers EGAS00001000374 and EGAS00001001306). Additional HCC data are available at GDC Data Portal (raw and processed data under project code TCGA-LIHC). CLL cell line RNA-Seq data are available at GSE134197.



## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. For genomic and clinical data analysis, we used all available data. For the rhAMP assay, we tested all but one CLL sample with RNA sequencing data (n=298). For RT-PCR, we randomly selected at least four samples per condition, which is sufficient for two-samples Wilcoxon test.
Data exclusions	To minimize false negatives, we excluded 103 HCC patients in the LIRI-JP project due to undetermined status for the U1 g.3A>C mutation. These samples do not have enough coverage for the U1 genes AND have no RNA-Seq data. These samples were excluded before any downstream splicing/expression/clinical analysis.
Replication	For the rhAMP assay, we used technical triplicates. One donor which showed inconsistent results for rhAMP was reported in the paper. For the RT-PCR, we used 6 U1 mutated samples and 4 wild-type samples for each CLL subtype as biological replicates; results were consistent. For qPCR, we used 6 U1 mutated samples and 4 wild-type samples for each CLL subtype as biological replicates; results were consistent. For cell line experiments, we used JVM3, HG3 and MEC1 as biological replicates and each condition (MUT or WT) was sequenced twice as technical replicates; results were consistent so we merged RNA-Seq reads from technical replicates.
Randomization	For differential expression and splicing analysis, we used randomized genotype labels (MUT or WT) as controls.
Blinding	Not relevant in this study as we studied the effect of a mutation from previously established patient cohorts.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	CLL cell lines HG3, MEC1 and JVM3 were obtained from DSMZ ( <a href="https://www.dsmz.de/catalogues/catalogue-human-and-animal-cell-lines.html">https://www.dsmz.de/catalogues/catalogue-human-and-animal-cell-lines.html</a> ).
Authentication	The authenticity of the cell lines was tested with the AmpFLSTR Identifier Plus PCR Amplification Kit.
Mycoplasma contamination	CLL cell lines have tested negative for mycoplasma.
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	No commonly misidentified lines were used.

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Median age at diagnosis: 60 years old (Standard Deviation = 19). Gender: 45% female / 55% male.
----------------------------	--

	<p>Diagnosis: 37 major cancer types. See Supplementary Table 1 for gender, age and diagnosis of individual patients.</p>
Recruitment	<p>No patients were recruited specifically for this project. All samples used in this study were from participants recruited by individual ICGC and TCGA projects.</p>
Ethics oversight	<p>Protocol approval is not required. For the PCAWG data, guidelines for study procedures were approved by the University of Toronto Research Ethics Board under RIS Human Protocol Number 30278 and protocol title "Pan-cancer Analysis of Whole Genomes: PAWG" . For CLL samples, guidelines for study procedures were approved by the Hospital Clinic of Barcelona Institutional Review Board under protocol number HCB/2015/0814 and protocol title "Functional and Clinical Impact of Genomic Analysis in CLL".</p>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Alcohol metabolism contributes to brain histone acetylation

<https://doi.org/10.1038/s41586-019-1700-7>

Received: 30 April 2018

Accepted: 17 September 2019

Published online: 23 October 2019

P. Mews<sup>1,2,4</sup>, G. Egervari<sup>1,4\*</sup>, R. Nativio<sup>1</sup>, S. Sidoli<sup>1,3</sup>, G. Donahue<sup>1</sup>, S. I. Lombroso<sup>1</sup>, D. C. Alexander<sup>1</sup>, S. L. Riesche<sup>1</sup>, E. A. Heller<sup>1</sup>, E. J. Nestler<sup>2</sup>, B. A. Garcia<sup>1</sup> & S. L. Berger<sup>1\*</sup>

Emerging evidence suggests that epigenetic regulation is dependent on metabolic state, and implicates specific metabolic factors in neural functions that drive behaviour<sup>1</sup>. In neurons, acetylation of histones relies on the metabolite acetyl-CoA, which is produced from acetate by chromatin-bound acetyl-CoA synthetase 2 (ACSS2)<sup>2</sup>. Notably, the breakdown of alcohol in the liver leads to a rapid increase in levels of blood acetate<sup>3</sup>, and alcohol is therefore a major source of acetate in the body. Histone acetylation in neurons may thus be under the influence of acetate that is derived from alcohol<sup>4</sup>, with potential effects on alcohol-induced gene expression in the brain, and on behaviour<sup>5</sup>. Here, using *in vivo* stable-isotope labelling in mice, we show that the metabolism of alcohol contributes to rapid acetylation of histones in the brain, and that this occurs in part through the direct deposition of acetyl groups that are derived from alcohol onto histones in an ACSS2-dependent manner. A similar direct deposition was observed when mice were injected with heavy-labelled acetate *in vivo*. In a pregnant mouse, exposure to labelled alcohol resulted in the incorporation of labelled acetyl groups into gestating fetal brains. In isolated primary hippocampal neurons *ex vivo*, extracellular acetate induced transcriptional programs related to learning and memory, which were sensitive to ACSS2 inhibition. We show that alcohol-related associative learning requires ACSS2 *in vivo*. These findings suggest that there is a direct link between alcohol metabolism and gene regulation, through the ACSS2-dependent acetylation of histones in the brain.

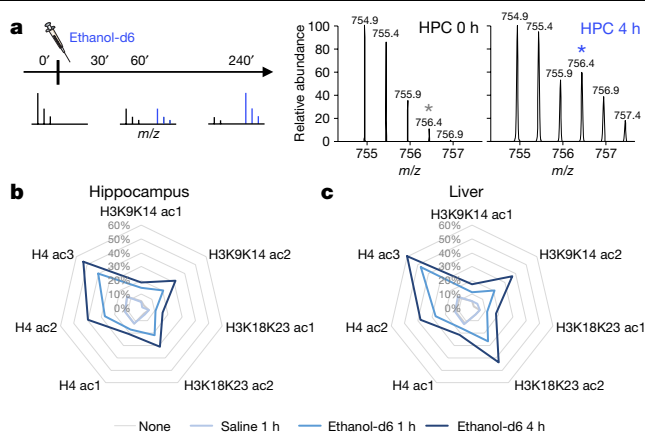
To determine whether acetate that is produced from the breakdown of alcohol contributes to dynamic acetylation of histones in the brain, we used *in vivo* stable-isotope labelling of protein acetylation, monitored by mass spectrometry<sup>6</sup> (Fig. 1a). Acetyl groups derived from ethanol were rapidly incorporated into histone acetylation in the brain, both in the hippocampus (Fig. 1b) and in the prefrontal cortex (Extended Data Fig. 1c). The incorporation of labelled alcohol into histone acetylation was dynamic in the hippocampus, and heavy labelling decreased to baseline levels 8 h after intraperitoneal injection (Extended Data Fig. 2a, b). This rapid *in vivo* labelling of histone acetylation in the hippocampus was absent in mice that were injected with non-labelled alcohol (Extended Data Fig. 2d). Similar labelling occurred in the liver (Fig. 1c), which is the principal site of alcohol metabolism and expresses high levels of ACSS2<sup>7,8</sup>. By contrast, the abundance of labelled acetylated histones was lower in skeletal muscle (gastrocnemius muscle; Extended Data Figs. 1d, 2e), which has relatively lower ACSS2 expression levels<sup>9</sup>.

To test whether ACSS2 is required for the incorporation of alcohol-derived acetate into the brain, we attenuated the expression of ACSS2 in the dorsal hippocampus (dHPC) by shRNA knockdown using a previously validated viral vector<sup>2</sup>. In these ACSS2-knockdown mice, we compared alcohol-derived histone acetylation separately in the dHPC, in which ACSS2 was reduced, and in the ventral hippocampus (vHPC),

in which ACSS2 expression was not affected. Notably, knockdown of ACSS2 prevented the incorporation of heavy-labelled acetyl groups derived from alcohol into histone acetylation in the dHPC (Fig. 2a). By contrast, in the same mouse, incorporation of the heavy label into histone acetylation in the vHPC was not affected (Fig. 2b). These *in vivo* data indicate that acetate that is derived from the metabolism of alcohol in the liver is transported to the brain and readily incorporated into ACSS2-dependent histone acetylation (see Supplementary Information for further discussion).

Although the majority of alcohol metabolism takes place in the liver, alcohol fractions may also be converted to acetate in the brain<sup>10</sup> by the enzymes catalase and cytochrome P450 2E1 (CYP2E1). We therefore assessed the contribution of acetyl groups that are derived from extracellular acetate to histone acetylation in the brain. In mice that were intraperitoneally injected with 2 g kg<sup>-1</sup> deuterated acetate (acetate-d<sub>3</sub>), the labelled acetate was rapidly incorporated into histone acetylation in the brain, at similar levels in both hippocampus and cortex (Extended Data Fig. 2f, g). Relative levels of the labelled acetate were highest at 30 min and returned to background levels at 4 h after injection, indicating that acetate-derived acetyl groups were quickly incorporated into brain histone acetylation and showing the rapid turnover of this process. Notably, we found that levels of acetate in the hippocampus

<sup>1</sup>Epigenetics Institute, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. <sup>2</sup>Fishberg Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>3</sup>Present address: Department of Biochemistry, Albert Einstein College of Medicine, New York, NY, USA. <sup>4</sup>These authors contributed equally: P. Mews, G. Egervari. \*e-mail: [egervari@penncmedicine.upenn.edu](mailto:egervari@penncmedicine.upenn.edu); [bergers@penncmedicine.upenn.edu](mailto:bergers@penncmedicine.upenn.edu)



**Fig. 1 | Alcohol metabolites contribute to histone acetylation in the brain.**

**a**, Experimental outline of in vivo ethanol- $d_6$  mass spectrometry. Asterisk indicates the isotopic shift owing to label incorporation, corresponding to the  $d_3$ -labelled form of the acetylated peptide. HPC, hippocampus. **b**, Metabolized heavy ethanol- $d_6$  is incorporated into histone acetylation in the hippocampus. The axis of the Arachne plot represents the percentage of the third isotope for the acetylated peptide, corresponding to the  $d_3$ -labelled form; the natural relative abundance of that isotope is apparent in the 'none' and 'saline 1 h' treatment groups. **c**, Label incorporation into histone acetylation occurs earlier in the liver, the principal site of alcohol metabolism.

were significantly increased 30 min after injection of alcohol or acetate (Extended Data Fig. 2h), and we detected substantial amounts of heavy acetate in the hippocampus as early as 30 min after injection with deuterated ethanol (ethanol- $d_6$ ) (Extended Data Fig. 3a).

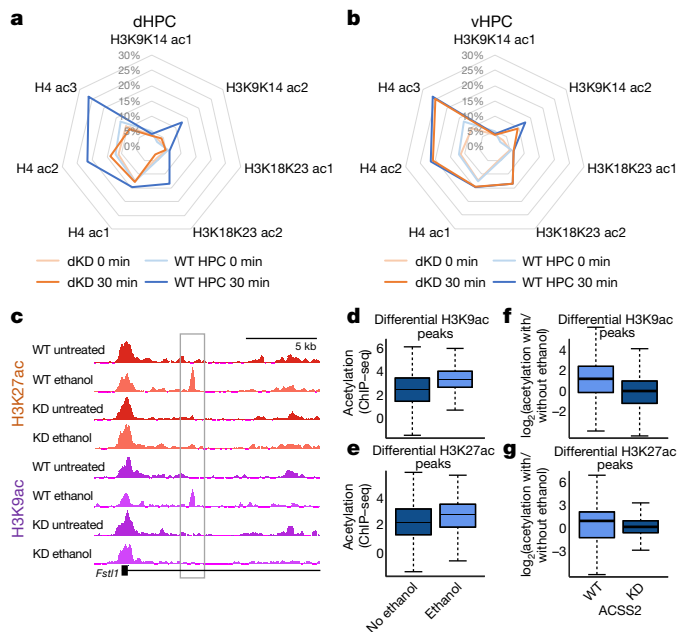
We further investigated whether alcohol-derived carbon groups are incorporated into other key metabolites in hippocampal tissue. Although we detected no incorporation of the alcohol-derived label into pools of glucose and 3-hydroxybutyrate, and only a fraction (less than 1%) into the pool of lactate, we found that alcohol labels pools of glutamine in the hippocampus (Extended Data Fig. 3b–e). In the brain, *de novo* synthesis of glutamine in astrocytes replenishes the glutamate–glutamine cycle, in which glutamine is trafficked into glutamatergic neurons for production of the neurotransmitter glutamate. Citrate—the substrate that is used by ATP citrate lyase to produce nucleo-cytoplasmic acetyl-CoA—is generated from  $\alpha$ -ketoglutarate, which can be derived from carboxylation of glutamine; this path could therefore provide another route through which alcohol contributes to histone acetylation. However, we detected only traces of the alcohol-derived label in pools of citrate or isocitrate in the hippocampus (Extended Data Fig. 3f). Taken together with our mass spectrometry data in ACSS2-knockdown mice (Fig. 2a, b), these results support the view that alcohol-derived acetate that contributes to the acetylation of histones in the hippocampus is converted to acetyl-CoA directly by ACSS2. Accordingly, our data suggest that the increased levels of blood acetate from alcohol metabolism promote ACSS2-mediated dynamic acetylation of histones in the brain.

We examined the functional relevance of alcohol-derived acetate for ACSS2-dependent histone acetylation in regulating the expression of genes in the hippocampus. We found that treating wild-type mice with alcohol resulted in significant enrichment of peaks for K9-acetylated histone H3 (H3K9ac) and H3K27ac—both at key neuronal genes and genome-wide—and that this enrichment was greatly attenuated in ACSS2-knockdown mice (Fig. 2c–g; chromatin immunoprecipitation followed by sequencing (ChIP-seq) performed 1 h after alcohol injection). For example, we observed ACSS2-dependent and alcohol-induced histone acetylation at *Fstl1* (follistatin-like 1) (Fig. 2c), a neuronal gene that has been implicated in the development and migration of neurons<sup>11</sup>, and alcohol-induced enrichment of H3K27ac at *Cep152* (centrosomal protein of 152 kDa) (Extended Data Fig. 4a), an important regulator of

genome integrity that is recurrently mutated in intellectual developmental disorders and microcephaly<sup>12</sup>. Another example is *Uimc1* (ubiquitin-interaction-motif-containing 1) (Extended Data Fig. 4b), which has previously been linked with neurodevelopmental disorders and autism<sup>13</sup>. Evaluating the ChIP-seq data for histone acetylation on a genome-wide scale, we found that 74% of the H3K9ac peaks that changed after exposure to alcohol were increased (339 out of 458 changed peaks called with MACS2; 10% false discovery rate (FDR) used as the significance threshold for DiffBind; Fig. 2d), and that 60% of the differential H3K27ac peaks were increased by alcohol (490 out of 816 peaks; ChIP-seq performed 1 h after alcohol injection; (Fig. 2e). Notably, this response was eliminated in ACSS2-knockdown mice: 98% of the H3K9ac and H3K27ac peaks that increased in wild-type mice after alcohol treatment were not induced in the dHPC of ACSS2-knockdown mice (Fig. 2f, g). We then performed RNA sequencing (RNA-seq) to characterize the transcriptional response and found that H3K9ac and H3K27ac drove gene expression genome-wide in wild-type mice that were treated with alcohol (Extended Data Fig. 5a, b). However, in line with the ChIP-seq data, this response was blunted in ACSS2-knockdown mice (Extended Data Fig. 5c, d). A functional analysis of the genes that were both hyperacetylated and induced by alcohol in an ACSS2-dependent manner revealed enrichment of genes with functions in protein binding, cell junctions, postsynaptic density and drug response (Extended Data Fig. 5e, f). Together, our in vivo findings show that treatment with alcohol leads to increased histone acetylation and transcriptional activity in the dHPC in an ACSS2-dependent manner.

Because alcohol and acetate have pleiotropic effects on brain circuitry and metabolism<sup>14</sup>, we developed an ex vivo assay to more closely model the direct effects of exogenous acetate on gene expression. We used isolated mouse primary hippocampal neurons, cultured for one week after isolation and subsequently treated with 5 mM acetate, to investigate the transcriptional response to supraphysiological levels of acetate that mimic the influx of exogenous acetate during alcohol intake. Furthermore, to determine the specific role of ACSS2 in transcriptional response to acetate, we used a highly specific small-molecule inhibitor<sup>2,7</sup> of ACSS2 (ACSS2i;  $C_{20}H_{18}N_4O_2S_2$ ) (Extended Data Fig. 6a).

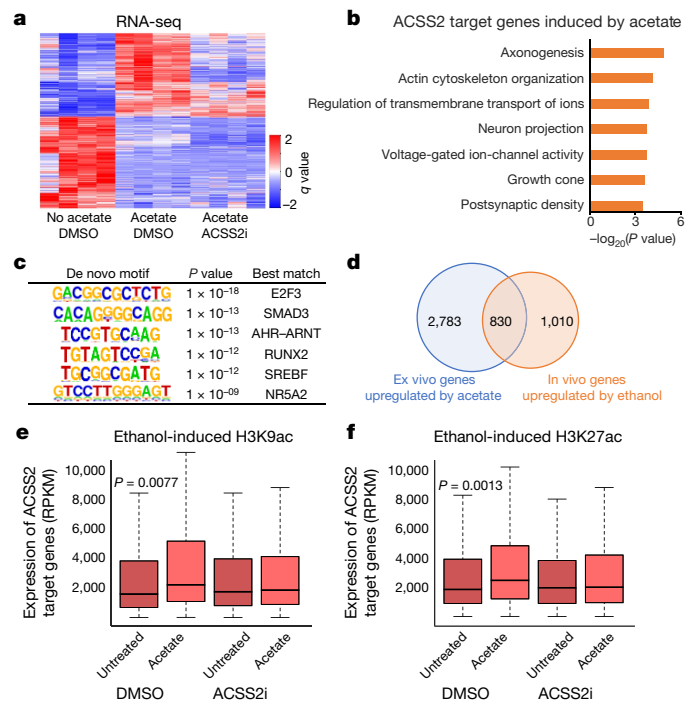
In primary hippocampal neurons, supplementation with acetate induced the expression of 3,613 genes (Fig. 3a, Extended Data Fig. 6b). Using an analysis of Gene Ontology (GO) terms, we found that these genes are involved in nervous system processes, including signal transduction and learning and memory (Extended Data Fig. 6c). By contrast, treatment with acetate resulted in the downregulation of genes that are involved in immune system processes (Extended Data Fig. 6d). In the presence of ACSS2i, 2,107 of the genes that were induced by acetate were no longer upregulated (Extended Data Fig. 6f), indicating that acetate-induced transcription relies heavily on the catalytic activity of ACSS2. Notably, acetate-induced genes were not regulated by treatment with ACSS2i in the absence of acetate (Extended Data Fig. 6e). GO analysis of the upregulated genes that were sensitive to ACSS2i showed enrichment for nervous system processes, behaviour, learning and memory (Extended Data Fig. 6f) and specific genes showed sensitivity to ACSS2i (Extended Data Fig. 7a–d). For example, *Slc17a7* was upregulated after treatment of wild-type hippocampus cells with acetate, but induction was diminished when ACSS2 was inhibited (Extended Data Fig. 7a). *Slc17a7* encodes vesicular glutamate receptor 1, which is implicated in hippocampal synaptic plasticity, addiction and alcohol use<sup>15</sup>. In addition, impaired DNA methylation of *Ccnj1* (cyclin J-like) has been linked to prenatal alcohol exposure and fetal alcohol spectrum disorder (FASD)<sup>16</sup> (Extended Data Fig. 7b). Further analysis revealed that the ACSS2i-sensitive and acetate-upregulated genes were also bound by hippocampal ACSS2 (which has been investigated by ChIP-seq previously<sup>2</sup>), and binding was proximal to the promoter at baseline without any direct behavioural stimulation in vivo<sup>2</sup> (Extended Data Fig. 8a). GO analysis linked these ACSS2 target genes to intricate plasticity-related mechanisms that involve axonogenesis and the activity of voltage-gated ion channels (Fig. 3b). Correspondingly, motif analysis of ACSS2-targeted, acetate-induced and



**Fig. 2 | Mass spectrometry analysis of ethanol-d6 in mice with knockdown of ACSS2 expression in the dHPC.** **a**, Knockdown of ACSS2 expression in the dHPC (dKD) prevents the incorporation of the heavy label into histone acetylation. WT, wild type. **b**, In the same mouse, incorporation of the heavy label into histone acetylation in the vHPC (where levels of ACSS2 are normal) is unchanged compared to control mice. **c**, ChIP-seq for H3K9ac and H3K27ac in untreated and ethanol-treated wild-type and ACSS2-knockdown (KD) mice ( $n = 3$  independent replicates). The genome-browser track view shows the *Fstl1* locus (chr16: 37,776,000–37,793,000). **d**, **e**, ChIP-seq for H3K9ac (**d**) and H3K27ac (**e**) in vivo shows increased acetylation genome-wide after injection of ethanol (339 out of 458 H3K9ac peaks and 490 out of 816 H3K27ac peaks called with MACS2; 10% FDR used as significance threshold for DiffBind).  $P < 2.2 \times 10^{-16}$  (**d**);  $P = 8.42 \times 10^{-11}$  (**e**) (two-sided Mann-Whitney rank-sum test). **f**, **g**, Induction of H3K9ac (**f**) and H3K27ac (**g**) is diminished in ACSS2-knockdown mice (458 H3K9ac peaks, 816 H3K27ac peaks).  $P < 2.2 \times 10^{-16}$  (**f**);  $P = 2.22 \times 10^{-6}$  (**g**) (two-sided Mann-Whitney rank-sum test). Box plots show the values for the first and third quartiles (box limits) and the median (centre), with whiskers extending to  $1.5 \times$  the interquartile range.

ACSS2i-sensitive genes implicated the involvement of neuronal transcription factors—including E2F3 and NR5A2 (Fig. 3c), which have been linked to neurodifferentiation and drug-related regulation of behaviour<sup>17,18</sup>.

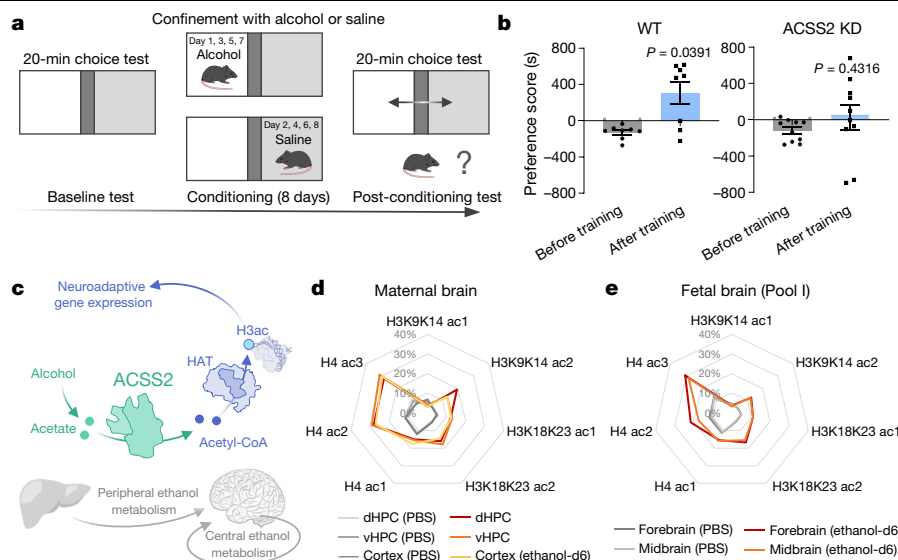
There was a substantial overlap between genes that were upregulated by alcohol in vivo in the dHPC and genes that were induced by acetate ex vivo—RNA-seq identified 830 alcohol-responsive hippocampal genes that overlapped with the ex vivo differentially expressed genes (Fig. 3d)—which suggests that translating our ex vivo model to the in vivo situation is valid. GO analysis for these overlapping genes highlighted the enrichment of genes that are related to neuronal plasticity, including those with roles in synapses, neuron projection and axons; however, genes related to ribosomal and mitochondrial functions were also enriched (Extended Data Fig. 8b). Notably, a previously published microarray dataset of hippocampal genes that are regulated by alcohol in vivo also showed substantial overlap with our list of genes that are induced by acetate ex vivo (81 of 214 (38%) alcohol-responsive hippocampal genes in the microarray<sup>19</sup>). Next, we showed (starting from a complementary analysis of our in vivo data) that target genes of ACSS2 in the hippocampus that show alcohol-induced H3K9ac in vivo were also upregulated by treatment of hippocampal neurons with acetate ex vivo, and that ACSS2i blocks this gene induction (Fig. 3e). The equivalent relationship existed for hippocampal genes that show alcohol-induced H3K27ac in vivo; these genes were not induced by acetate ex vivo in the presence of ACSS2i (Fig. 3f).



**Fig. 3 | Acetate-induced transcription mediated by ACSS2 in primary hippocampal neurons.** **a**, RNA-seq in primary hippocampal neurons that were isolated from C57/BL6 mouse embryos and treated with acetate (5 mM) in the presence of ACSS2i or vehicle (DMSO). The first two columns of the heat map show 7,600 genes that were differentially expressed after treatment with acetate, and the third column shows the behaviour of those genes in the presence of ACSS2i. A total of 2,107 of the 3,613 genes that were induced by acetate were not upregulated in the presence of ACSS2i ( $n = 4$  per group). **b**, GO analysis of genes that are both sensitive to acetate and directly bound by ACSS2. Data are from ACSS2 ChIP-seq ( $n = 429$  genes). Genes were assessed using a modified Fisher's exact test (EASE) with the FDR corrected by the Yekutieli procedure;  $-\log_{10}$  transformations of nominal  $P$  values are shown. **c**, HOMER (hypergeometric optimization of motif enrichment) unsupervised analysis of hippocampal ACSS2-binding sites in the hippocampus, targeting acetate-sensitive genes. De novo motif analysis of 751 ACSS2 peaks; a hypergeometric test was used for each motif to compare to the background set of ACSS2 peaks that do not target acetate-sensitive genes. **d**, Overlap of genes that are upregulated by treatment with ethanol in vivo (dHPC) and acetate in vitro ( $n = 830$ ).  $P = 3.48 \times 10^{-237}$  (hypergeometric test of gene-set overlap). **e**, **f**, ACSS2 target genes with alcohol-induced H3K9ac (**e**) and H3K27ac (**f**) in vivo are upregulated by acetate in hippocampal neurons in vitro. ACSS2i blocks this gene induction.  $n = 285$  genes (**e**) and  $n = 362$  genes (**f**), tested against an equal number of control genes using a two-sided Mann-Whitney rank-sum test;  $P = 0.0077$  (**e**),  $P = 0.0013$  (**f**). Gene expression is given in reads per kilobase of transcript per million mapped reads (RPKM). Box plots as in Fig. 2.

Together, these findings suggest that ACSS2 may have a role in alcohol-related learning by coordinating alcohol-induced histone acetylation and gene expression. To examine potential behavioural effects in wild-type and ACSS2-knockdown mice, we performed ethanol-mediated conditioned place preference (CPP), which has previously been used to assess ethanol-associated learning<sup>20</sup>. In this paradigm, mice are exposed to neutral and rewarding stimuli in distinct spatial compartments, distinguished by environmental cues. After conditioning, CPP is measured by allowing the mice free access to either compartment and measuring the time spent in the chamber that is associated with rewarding stimuli (Fig. 4a). To assess place-preference learning, we calculated the mean time spent in the conditioned (paired with alcohol) and unconditioned (paired with saline) chambers (Extended Data Fig. 9c), as well as a CPP score, which we defined as the difference between the time spent in the conditioned and the unconditioned chamber (Fig. 4b). We found that





**Fig. 4 | ACSS2 is required for alcohol-induced associative learning.**

**a**, Schematic of ethanol-induced CPP. **b**, Preference scores for the ethanol-paired chamber in wild-type mice ( $n = 8$ ;  $P = 0.0391$ ) and mice with knockdown of ACSS2 in the dHPC ( $n = 10$ ;  $P = 0.4316$ ). Data are mean  $\pm$  s.e.m.;  $P$  values by Wilcoxon matched-pairs signed-rank test. **c**, Model. Acetate from the breakdown of alcohol in the liver is activated by neuronal ACSS2 in the brain and induces gene-regulatory acetylation of histones. HAT, histone acetyltransferase.

**d**, Metabolized ethanol-d6 is incorporated into histone acetylation in the maternal brain. **e**, Incorporation of ethanol-d6 into histone acetylation in the fetal brain. Data represent the first of two pools of embryos ( $n = 4$  per pool) from maternal ethanol-d6 injection. The axes of the Arachne plots represent the percentage of the third isotope for the acetylated peptide, corresponding to the  $d_3$ -labelled form.

wild-type mice spent longer in the compartment to which ethanol was delivered during training ( $P = 0.0391$ ; Wilcoxon test) (Fig. 4b). Notably, acquisition of CPP depends on spatial memory formation in the dHPC and, accordingly, dHPC lesions disrupt place conditioning<sup>21</sup>. To test the importance of ACSS2 in the dHPC, we injected mice with AAV9 virus expressing shRNA against *Acss2* to reduce the protein level of ACSS2 ( $n = 10$ ), or control virus expressing GFP only ( $n = 8$ ) (Extended Data Fig. 9a, b). We detected a significant main effect of the conditioning subgroup ( $P = 0.0227$ ,  $F_{1,32} = 5.731$ ; main effect of training from a two-way analysis of variance (ANOVA) across the four groups), showing that the ethanol-induced CPP procedure was successful. In addition, we observed a significant interaction between the ACSS2-knockdown and conditioning subgroups ( $P = 0.0462$ ,  $F_{1,32} = 4.303$ ; interaction from two-way ANOVA across the four groups), indicating that the knockdown of ACSS2 in the dHPC significantly reduced the expression of CPP. Strikingly, we found that ethanol-associated CPP was suppressed in mice in which ACSS2 expression was reduced in the dHPC ( $P = 0.4316$ ; Wilcoxon test) (Fig. 4b), indicating that the formation of ethanol-related associative memories requires ACSS2.

Overall, our ex vivo and in vivo molecular data, together with our behavioural findings, show that ACSS2 is required for the incorporation of heavy-labelled acetate into acetylated histones in the dHPC, and that this facilitates memory-related gene expression and alcohol-related associative learning (Fig. 4c). These results establish ACSS2 as a promising candidate for therapeutic intervention in alcohol-use disorders, in which the memory of alcohol-associated environmental cues is a primary driver of craving and relapse.

Alcohol exposure not only disrupts epigenetic and transcriptional processes in the adult brain, but is also linked to epigenetic dysregulation in the gestating fetus<sup>22</sup>. In utero, alcohol is an environmental teratogen that affects the expression of neurodevelopmental genes and can elicit numerous alcohol-associated postnatal disease phenotypes, which together are categorized as FASD<sup>23</sup>. Previous investigations of alcohol-mediated epigenetic changes in utero have implicated altered histone acetylation in FASD<sup>24</sup>, but the underlying mechanisms are unknown.

We investigated whether alcohol affects dynamic histone acetylation in utero in the developing fetal midbrain and forebrain of mice at embryonic day (E)18.5. Mass spectrometry of fetal brain showed that 'binge-drinking-like' alcohol exposure, in parallel with maternal labelling of neuronal histone acetylation, resulted in the deposition of alcohol-derived acetyl groups onto histones in the fetal forebrain and midbrain in early neural development (Fig. 4e, Extended Data Fig. 9d), indicating an unanticipated potential mechanism for the aetiology of FASD.

In the adult brain, epigenetic mechanisms that control gene expression have a key role in processing neural activity to continuously adapt circuit connectivity and behaviour<sup>25,26</sup>. Here, we show that exposure to alcohol gives rise to the acetylation of histones in the brain both directly (through the direct incorporation of alcohol-derived acetate) and indirectly (through other metabolic pathways). Incorporation of alcohol-derived acetate into histone acetylation was recently observed in the liver<sup>27</sup>. However, to our knowledge, our data provide the first empirical evidence indicating that a portion of acetate that is derived from the metabolism of alcohol directly influences epigenetic regulation in the brain. We show that this direct pathway has important functional and behavioural consequences, shedding light on a neurobiological aspect of alcohol use. Given that effects of ethanol on the brain and behaviour are complex, further studies will be required to determine the relative contributions of ethanol-derived histone acetylation, ethanol-induced intracellular signalling pathways and ethanol-related redox stress. We also show that histone acetylation in the brain occurs through the generation of acetyl-CoA by ACSS2. In the hippocampus, the incorporation of acetyl groups that are derived from alcohol may be critical for alcohol-related associative learning, which encodes environmental cues associated with alcohol that drive craving, seeking and consumption even after protracted periods of abstinence. The direct pathway that we identify here substantially furthers our understanding of alcohol-induced epigenetic regulation in the brain, which has previously been limited to the indirect effects of alcohol-induced intracellular signalling and changes in the expression or activity of histone-modifying enzymes. The direct pathway contributes to a large proportion of the histone acetylation that occurs after ethanol exposure, and suggests that the incorporation of

alcohol-derived acetyl groups is physiologically relevant and associated with the transcriptional and behavioural adaptations that are induced by ethanol. Notably, our findings suggest that other peripheral sources of physiological acetate—primarily the gut microbiome—may affect central histone acetylation and brain function in a similar manner, which may either control or foster other metabolic syndromes. Translational treatment strategies that target this nexus between peripheral metabolic activity and neuroepigenetic regulation may pave the way for therapeutic interventions for alcohol use and other neuropsychiatric disorders.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1700-7>.

- Li, X., Egervari, G., Wang, Y., Berger, S. L. & Lu, Z. Regulation of chromatin and gene expression by metabolic enzymes and metabolites. *Nat. Rev. Mol. Cell Biol.* **19**, 563–578 (2018).
- Mews, P. et al. Acetyl-CoA synthetase regulates histone acetylation and hippocampal memory. *Nature* **546**, 381–386 (2017).
- Sarkola, T., Iles, M. R., Kohlenberg-Mueller, K. & Eriksson, C. J. P. Ethanol, acetaldehyde, acetate, and lactate levels after alcohol intake in white men and women: effect of 4-methylpyrazole. *Alcohol. Clin. Exp. Res.* **26**, 239–245 (2002).
- Soliman, M. L. & Rosenberger, T. A. Acetate supplementation increases brain histone acetylation and inhibits histone deacetylase activity and expression. *Mol. Cell. Biochem.* **352**, 173–180 (2011).
- Pandey, S. C., Kyzar, E. J. & Zhang, H. Epigenetic basis of the dark side of alcohol addiction. *Neuropharmacology* **122**, 74–84 (2017).
- Mews, P. & Berger, S. L. in *Methods in Enzymology* Vol. 574 (ed. Marmorstein, R.) 311–329 (Elsevier, 2016).
- Comerford, S. A. et al. Acetate dependence of tumors. *Cell* **159**, 1591–1602 (2014).
- Zakhari, S. Alcohol metabolism and epigenetics changes. *Alcohol Res.* **35**, 6–16 (2013).
- Bonthuis, P. J. et al. Noncanonical genomic imprinting effects in offspring. *Cell Rep.* **12**, 979–991 (2015).
- Zimatkin, S. M., Pronko, S. P., Vasilou, V., Gonzalez, F. J. & Deitrich, R. A. Enzymatic mechanisms of ethanol oxidation in the brain. *Alcohol. Clin. Exp. Res.* **30**, 1500–1505 (2006).
- Liu, R. et al. *Fstl1* is involved in the regulation of radial glial scaffold development. *Mol. Brain* **8**, 53 (2015).
- Kalay, E. et al. CEP152 is a genome maintenance protein disrupted in Seckel syndrome. *Nat. Genet.* **43**, 23–26 (2011).
- Stessman, H. A. F. et al. Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat. Genet.* **49**, 515–526 (2017).
- Volkow, N. D. et al. Acute alcohol intoxication decreases glucose metabolism but increases acetate uptake in the human brain. *Neuroimage* **64**, 277–283 (2013).
- Rao, P. S. S., Bell, R. L., Engleman, E. A. & Sari, Y. Targeting glutamate uptake to treat alcohol use disorders. *Front. Neurosci.* **9**, 144 (2015).
- Laufer, B. I. et al. Associative DNA methylation changes in children with prenatal alcohol exposure. *Epigenomics* **7**, 1259–1274 (2015).
- Cates, H. M. et al. Transcription factor E2F3a in nucleus accumbens affects cocaine action via transcription and alternative splicing. *Biol. Psychiatry* **84**, 167–179 (2018).
- Stergiopoulos, A. & Politis, P. K. Nuclear receptor NR5A2 controls neural stem cell fate decisions during development. *Nat. Commun.* **7**, 12230 (2016).
- Mulligan, M. K. et al. Molecular profiles of drinking alcohol to intoxication in C57BL/6J mice. *Alcohol. Clin. Exp. Res.* **35**, 659–670 (2011).
- Juarez, B. I. et al. Midbrain circuit regulation of individual alcohol drinking behaviors in mice. *Nat. Commun.* **8**, 2220 (2017).
- Ferbinteanu, J. & McDonald, R. J. Dorsal/ventral hippocampus, fornix, and conditioned place preference. *Hippocampus* **11**, 187–200 (2001).
- Veazey, K. J., Parnell, S. E., Miranda, R. C. & Golding, M. C. Dose-dependent alcohol-induced alterations in chromatin structure persist beyond the window of exposure and correlate with fetal alcohol syndrome birth defects. *Epigenetics Chromatin* **8**, 39 (2015).
- Mead, E. A. & Sarkar, D. K. Fetal alcohol spectrum disorders and their transmission through genetic and epigenetic mechanisms. *Front. Genet.* **5**, 154 (2014).
- Mandal, C., Halder, D., Jung, K. H. & Chai, Y. G. In utero alcohol exposure and the alteration of histone marks in the developing fetus: an epigenetic phenomenon of maternal drinking. *Int. J. Biol. Sci.* **13**, 1100–1108 (2017).
- Mews, P. & Calipari, E. S. in *Progress in Brain Research* Vol. 235 (eds Calvey, T. & Daniels, W.) 19–63 (Elsevier, 2017).
- Egervari, G., Ciccocioppo, R., Jentsch, J. D. & Hurd, Y. L. Shaping vulnerability to addiction - the contribution of behavior, neural circuits and molecular mechanisms. *Neurosci. Biobehav. Rev.* **85**, 117–125 (2018).
- Kriss, C. L. et al. In vivo metabolic tracing demonstrates the site-specific contribution of hepatic ethanol metabolism to histone acetylation. *Alcohol. Clin. Exp. Res.* **42**, 1909–1923 (2018).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## Methods

### Data reporting

No statistical methods were used to predetermine sample size. All animals were randomly allocated into experimental and control groups. Mass spectrometry analysis of heavy labelled samples and analysis of behavioural experiments was performed by blinded investigators.

### Stable isotope labelling of brain histone acetylation

We injected mice intraperitoneally with 2 g kg<sup>-1</sup> ethanol-d6 or control saline, and assessed deuterium incorporation into acetylated histones at baseline, as well as at 1 and 4 h after intraperitoneal injections (Fig. 1a). We confirmed that the injected ethanol-d6 is readily metabolized to acetate that becomes systemically accessible (Extended Data Fig. 1a), resulting in rapid labelling of blood acetate (Extended Data Fig. 1b). Using quantitative liquid chromatography–mass spectrometry (LC–MS), we quantified the relative abundance of isotopically labelled histone acetylation in the brain and in peripheral tissues (Fig. 1a). Analysis of the patterns of heavy labelling showed that injection of heavy-labelled ethanol leads to marked increases of the M+3 species (Extended Data Fig. 1e, f, Supplementary Table 1), which suggests that ethanol-derived acetate contributes to histone acetylation through direct deposition of triply deuterated acetyl-CoA. This increase in the M+3 species was also evident when accounting for the natural levels of heavy isotopes of carbon and hydrogen (Extended Data Fig. 1g, h; see also Methods for detailed descriptions). The increases of M+1 and M+2 species (Extended Data Fig. 1h) could be driven by deuterium back exchange<sup>28</sup> or by singly and doubly deuterated metabolites, indicating that alternative metabolic pathways also contribute to histone acetylation labelling; however, the major increase is to the M+3 species through triply deuterated acetate. We also performed tracing of <sup>13</sup>C-labelled alcohol and found rapid incorporation into histone acetylation in the hippocampus, equivalent to ethanol-d6 (Extended Data Fig. 2c). We examined whether alcohol exposure during gestation influences histone acetylation in the developing fetal brain by measuring the direct deposition of alcohol-derived acetyl groups onto histones. Using the protocol described above (intraperitoneal injection of heavy-labelled alcohol (2 g kg<sup>-1</sup>) followed by mass spectrometry on isolated histone proteins) we confirmed incorporation of alcohol metabolites (4 h after injection) into the acetylation of histones in neurons of gestating female mice (Fig. 4d), consistent with the previous results in males (Fig. 1b).

### Histone extraction

Histones were extracted as previously described<sup>2</sup>. The cells were incubated in nuclear isolation buffer (15 mM Tris-HCl, 15 mM NaCl, 60 mM KCl, 5 mM MgCl<sub>2</sub>, 1 mM CaCl<sub>2</sub> and 250 mM sucrose at pH 7.5; 0.5 mM AEBSF, 10 mM sodium butyrate, 5 nM microcystin and 1 mM DTT added fresh) with 0.2% NP-40 on ice for 5 min. The nuclei were collected by centrifuging at 700g at 4 °C for 5 min. The resulting nuclear pellet was washed twice with the same volume of nuclear isolation buffer without NP-40. Histones were then acid-extracted with 0.2 M H<sub>2</sub>SO<sub>4</sub> for 3 h at 4 °C with rotation. The insoluble nuclear debris was pelleted at 3,400g at 4 °C for 5 min, and the supernatant was retained. Next, histone proteins were precipitated by adding 100% trichloroacetic acid in a 1:3 ratio (v/v) for 1 h at 4 °C. The pellet was washed with acetone to remove residual acid. Histones were resuspended in 30 µl of 50 mM NH<sub>4</sub>HCO<sub>3</sub> (pH 8.0).

### Histone propionylation and digestion

Histones were derivatized and digested as previously described<sup>6</sup>. The sample was mixed with 15 µl derivatization mix, consisting of propionic anhydride and acetonitrile in a 1:3 ratio (v/v), and this was immediately followed by the addition of 7.5 µl ammonium hydroxide to maintain pH 8.0. The sample was incubated for 15 min at 37 °C, dried and the

derivatization procedure was repeated one more time. Samples were then resuspended in 50 mM NH<sub>4</sub>HCO<sub>3</sub> and incubated with trypsin (enzyme:sample ratio of 1:20) overnight at room temperature. After digestion, the derivatization reaction was performed again twice to derivatize the N termini of the peptides. Samples were desalted using C18 stage tips before LC–MS analysis.

### Nanoscale liquid chromatography–tandem mass spectrometry

Samples were analysed by using a nanoscale liquid chromatography–tandem mass spectrometry (LC–MS/MS) setup. Nanoscale liquid chromatography was configured with a 75 µm ID × 25 cm Reprosil-Pur C18-AQ nano-column (3 µm; Dr. Maisch HPLC GmbH) using an EASY-nLC nano-flow high-performance liquid chromatography (HPLC) system (Thermo Fisher Scientific), packed in-house. The HPLC gradient was as follows: 5% to 32% solvent B in solvent A (A = 0.1% formic acid; B = 80% acetonitrile, 0.1% formic acid) over 45 min, from 32% to 90% solvent B in 5 min and 90% B for 10 min at a flow rate of 300 nl min<sup>-1</sup>. Nanoscale liquid chromatography was coupled to an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific). The acquisition method was data-independent acquisition (DIA) as described<sup>29</sup>. In brief, two full-scan mass spectra (*m/z* 300–1,100) were acquired in the ion trap within a DIA duty cycle, and 16 MS/MS were performed with an isolation window of 50 Da. Normalized collision energy was set to 35%.

### Data analysis

Raw mass spectrometry data were analysed manually. We selected the seven most-intense peptides of histones H3 and H4 that contained acetylations, and extracted the relative abundance of the M+1, M+2 and M+3 signals compared to the monoisotopic signal. The other peptides were not considered as, owing to their low abundance, we could not reliably quantify the relative abundance of all the isotopes. The percentage represented in the radar plots indicates the relative intensity of the M+3 signal (the fourth isotope) as compared to the monoisotopic signal. Data were not normalized to the non-labelled sample, so that the relative abundance of the natural isotopic distribution could also be observed in the untreated mice.

For statistical analysis, we performed two-tailed heteroscedastic *t*-tests (significant at *P* < 0.05) when comparing the same isotope in treated versus untreated samples (Supplementary Table 1). In this analysis, we considered differences in the relative abundance of all species (M+1, M+2, M+3), and found that major changes occur in M+3 (Supplementary Tables 1, 2). We used the R package *enviPat*<sup>30</sup> to estimate the theoretical relative abundance of the first four isotopes of the seven peptides considered in this study, which showed no significant difference to the observed isotopic distribution of the untreated samples (Extended Data Fig. 1g). Natural abundance corrections were performed using *FluxFix*<sup>31</sup>, to calculate the exact relative abundance of the M+1, M+2 and M+3 species in the samples that were treated with labelled ethanol or acetate (Extended Data Fig. 1h). We used a matrix-based approach to correct natural abundance, as proposed previously<sup>32</sup>. This calculation corrects for the contribution of each isotopologue, removing, for example, from the M+3 signal, the portion of the isotopic pattern that is contributed by the increase of the M+1 and M+2 species.

### ChIP-seq

ChIP-seq was performed as previously described with modifications for the preparation of mouse brains<sup>33</sup>. In brief, approximately 20 mg of dHPC from each mouse was minced on ice, cross-linked with 1% formaldehyde for 10 min and quenched with 125 mM glycine for 5 min. Nuclei were prepared by dounce homogenization of cross-linked tissue in nuclei isolation buffer (50 mM Tris-HCl at pH 7.5, 25 mM KCl, 5 mM MgCl<sub>2</sub>, 0.25 M sucrose) with freshly added protease inhibitors and sodium butyrate. Nuclei were lysed in nuclei lysis buffer (10 mM

Tris-HCl at pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% sodium deoxycholate, 0.5% *N*-lauroylsarcosine) with freshly added protease inhibitors and sodium butyrate, and chromatin was sheared using a Covaris S220 sonicator to approximately 250 bp in size. Equal aliquots of sonicated chromatin were used per immunoprecipitation reaction with 5  $\mu$ l H3K9ac antibody (Active Motif; 39137; 09811002) or 4  $\mu$ l H3K27ac antibody (Abcam; 4729, GR323132-1) pre-conjugated to Protein G Dynabeads (Life Technologies). Ten percent of the chromatin was saved as input DNA. ChIP reactions were incubated overnight at 4 °C with rotation and washed three times in wash buffer. Immunoprecipitated DNA was eluted from the beads, reversed cross-linked and purified together with the input DNA. Exactly 10 ng DNA (either ChIP or input) was used to construct sequencing libraries using the NEBNext Ultra II DNA library preparation kit for Illumina (New England Biolabs; NEB). Libraries were multiplexed using NEBNext Multiplex Oligos for Illumina (dual index primers) and single-end sequenced (75 bp) on the NextSeq 500 platform (Illumina) in accordance with the manufacturer's protocol.

### ChIP-seq analysis

ChIP-seq tags generated with the NextSeq 500 platform were demultiplexed with the bcl2fastq utility and aligned to the mouse reference genome (assembly GRCm38 (mm10)) using Bowtie v.1.1.1, allowing up to two mismatches per sequencing tag<sup>34</sup> (parameters -m 1 -best). Peaks were detected using MACS2 (tag size = 75 bp; FDR < 1  $\times$  10<sup>-3</sup>) from pooled H3K9ac or H3K27ac tags of mice from the same condition along with treatment-matched input tags as control. The multiple transcription-factor-binding loci (MTL) method<sup>35</sup> was used to compare H3K9ac or H3K27ac enrichment in the four study conditions. The statistical significance of differential H3K9ac or H3K27ac enrichments was assessed by using DiffBind (Bioconductor v.3.7) in a two-way comparison (wild-type mice treated with saline versus wild-type mice treated with ethanol, or ACSS2-knockdown mice treated with saline versus ACSS2-knockdown mice treated with ethanol) across the individual replicate samples (FDR < 10%). UCSC genome-browser track views were created for ChIP-seq data by first pooling replicates and generating coverage maps using BEDtools genomeCoverageBed -bg, then adjusting for library size using the RPM (reads per million) coefficient. The input signal was then subtracted from the ChIP signal. The resulting tracks were converted from bedGraph to bigWig files using the bedGraphToBigWig function in the UCSC genome browser. RNA-seq tracks were created similarly, first splitting by tag orientation to the genomic reference strand and then creating coverage maps. Because an RPM adjustment might disguise a large deformation in the transcriptome distribution, maps were adjusted for library size using the average scalar coefficient size factor determined by DESeq2. The resulting tracks were converted to bigWig files, as for ChIP-seq tracks, and the + and - tags from a given sample were plotted as overlays in a track hub.

### RNA-seq

All RNA-seq data were prepared for analysis as follows. NextSeq sequencing data were demultiplexed using native applications on BaseSpace. Demultiplexed FASTQs were aligned by RNA-STAR v.2.5.2 to the assembly mm10 (GRCm38). Aligned reads were mapped to genomic features using HTSeq v.0.6.1. Quantification, library size adjustment and analysis of differential gene expression was done using DESeq2 and Wald's test (pairwise comparisons between treatment with acetate and DMSO in the inhibitor-treated or untreated cells, followed by a set overlap of differentially expressed genes). Overlaps between lists of genes were tested for significance using a hypergeometric test.

### Functional analysis

GO analysis was performed using DAVID with an FDR cut-off of 10%, filtering categories with fewer than 10 genes or less than 2.5 $\times$  fold enrichment

over background. Motif analysis was performed using HOMER v.4.6 on all ACSS2 peaks from published *in vivo* data<sup>2</sup>, targeting (by the nearest transcription start site) a gene sensitive to acetate with H3K9ac at the promoter using a fixed window of 300 bp<sup>2</sup>.

### Primary hippocampal neurons

Plated primary hippocampal neurons were obtained from the Neurons R Us neuronal core at the University of Pennsylvania. Cells were maintained in neurobasal medium (NBM; Gibco) supplemented with GlutaMAX (Gibco) and B27 (Gibco). After 7 days of differentiation, cells were treated for 24 h with 5 mM acetate or vehicle (NBM) in the presence or absence of 20  $\mu$ M ACSS2i or vehicle (DMSO diluted into NBM). Cells were lysed using QIAzol (Qiagen) and RNA was extracted using the RNeasy Mini Kit (Qiagen). RNA-seq libraries were prepared using the NEBNext Ultra II Directional RNA library preparation kit (NEB).

### Mouse experiments

Animal use, surgical procedures and all experiments performed were approved by the Institutional Animal Care and Use Committee (protocol 804849). All personnel involved have been adequately trained and are qualified according to the Animal Welfare Act and the Public Health Service policy. Adult male mice (8 weeks old) or E18.5 pregnant females were used. Ethanol, ethanol-d6 (Sigma-Aldrich; 186414), ethanol-1-<sup>13</sup>C (Sigma-Aldrich; 324523) and sodium acetate-d3 (Sigma-Aldrich) were administered by intraperitoneal injection and dosed at 2 g kg<sup>-1</sup> (20% solution in saline, filtered through a Stericup GV filter). CPP was performed as described previously<sup>36</sup>. In brief, mouse CPP boxes (Ugo Basile; 42553) with external dimensions 35  $\times$  18  $\times$  29 cm were used. The apparatus was divided into two chambers (16  $\times$  15  $\times$  25 cm) that differed in wall and floor pattern. Striped walls were paired with circle cutouts (1 cm) and solid grey walls were paired with square cutouts (0.5 cm). Sessions were run in a dark room at ambient temperature. Boxes were cleaned with 70% ethanol between mice and allowed to dry between rounds. The paradigm consisted of 1 habituation day (5 min exploration in neutral environment), 1 pre-training session (20 min with access to both chambers), 8 training days (biased subject assignment, alternating sessions of intraperitoneal injection of saline or ethanol immediately before the 5-min session) and 1 post-training test session (20 min with access to both chambers). The percentage of time spent in the conditioned chamber was measured by blinded investigators. Preference scores were calculated as the difference between the time spent in the conditioned chamber and the unconditioned chamber. A Shapiro-Wilk test was used to assess normal distribution and a Mann-Whitney test to determine learning.

### Intracranial injection of viral vector

Adult male C57BL/6J mice of 10 weeks of age were anaesthetized with isoflurane gas (1–5% to maintain surgical plane) and placed in a sterile field within a stereotaxic device. Artificial tears were applied to eyes to ensure sufficient lubrication. Mice received an injection of bupivacaine (2.5 mg kg<sup>-1</sup>) for local anaesthesia before the skin was disinfected with betadine solution and the skull exposed with a short incision using sterile surgical equipment. A small hole (about 0.5 mm) was drilled in the skull over the target area using a stereotax and a stereotactic drill. A micro-syringe filled with viral vector was inserted into the dHPC and slowly removed following injection (AP, -2.0 mm; DV, -1.4 mm; ML,  $\pm$  1.5 mm from bregma). The vector for ACSS2 knockdown was AAV2/9.U6.shACSS2.CMV.EGFP. All mice received a single dose of subcutaneous meloxicam (5 mg kg<sup>-1</sup>) as analgesia at induction and one dose per day for two days postoperatively as needed.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

All RNA-seq and ChIP-seq data are available at the Gene Expression Omnibus (GEO) with accession number GSE122188. Raw mass spectrometry data are provided as Supplementary Table 2.

28. Linderstrom-Lang, K. Deuterium exchange between peptides and water. *Chem. Soc. Spec. Publ.* **2**, 1–20 (1955).
29. Sidoli, S., Simithy, J., Karch, K. R., Kulej, K. & Garcia, B. A. Low resolution data-independent acquisition in an LTQ-Orbitrap allows for simplified and fully untargeted analysis of histone modifications. *Anal. Chem.* **87**, 11448–11454 (2015).
30. Loos, M., Gerber, C., Corona, F., Hollender, J. & Singer, H. Accelerated isotope fine structure calculation using pruned transition trees. *Anal. Chem.* **87**, 5738–5744 (2015).
31. Trefely, S., Ashwell, P. & Snyder, N. W. FluxFix: automatic isotopologue normalization for metabolic tracer analysis. *BMC Bioinformatics* **17**, 485 (2016).
32. Feroandez, C. A., Rosiers, C. Des, Previs, S. F., David, F. & Brunengraber, H. Correction of <sup>13</sup>C mass isotopomer distributions for natural stable isotope abundance. *J. Mass Spectrom.* **31**, 255–262 (1996).
33. Nativio, R. et al. Dysregulation of the epigenetic landscape of normal aging in Alzheimer's disease. *Nat. Neurosci.* **21**, 497–505 (2018).
34. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
35. Chen, X. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
36. Cunningham, C. L., Gremel, C. M. & Groblewski, P. A. Drug-induced conditioned place preference and aversion in mice. *Nat. Protocols* **1**, 1662–1670 (2006).

**Acknowledgements** We thank the Metabolomics Core of the Diabetes Research Center (DRC) for providing the mass spectrometry quantification of metabolites; J. D. Rabinowitz (the Princeton Metabolomics Core director) and C. Jang for advice; and the Neurons R Us core of the Mahoney Institute for Neurological Sciences for preparations of primary hippocampal neurons. We especially acknowledge J. Whetstone for the suggestion to test whether the administration of alcohol to a pregnant female mouse leads to histone acetylation in the gestating fetal brain. G.E. was supported by The Brody Family Medical Trust Fund Fellowship in Incurable Diseases of The Philadelphia Foundation. This work was supported by NIH P01AG031862 and NIH R01AA027202.

**Author contributions** P.M. and S.L.B. developed the primary hypothesis; P.M., G.E. and S.L.B. designed the project; P.M. and G.E. performed most of the experiments; P.M. planned the ethanol-d6-labelling mass spectrometry, which was performed together with R.N., G.E. and S.S. with support from B.A.G.; G.E. and S.S. performed the acetate-d3-labelling mass spectrometry; the RNA-seq was performed in vivo by P.M. and in vitro by G.E., with support from R.N.; R.N. conducted the ChIP-seq; G.D. analysed all ChIP-seq and RNA-seq datasets; G.E., S.I.L. and D.C.A. performed the behavioural characterization with support from E.A.H.; P.M. performed the labelling experiments for metabolomic analysis with support from E.J.N.; the fetal alcohol labelling was done by G.E., with support from S.L.R.; P.M., G.E. and S.L.B. wrote the manuscript. All authors reviewed the manuscript and discussed the work.

**Competing interests** The authors declare no competing interests.

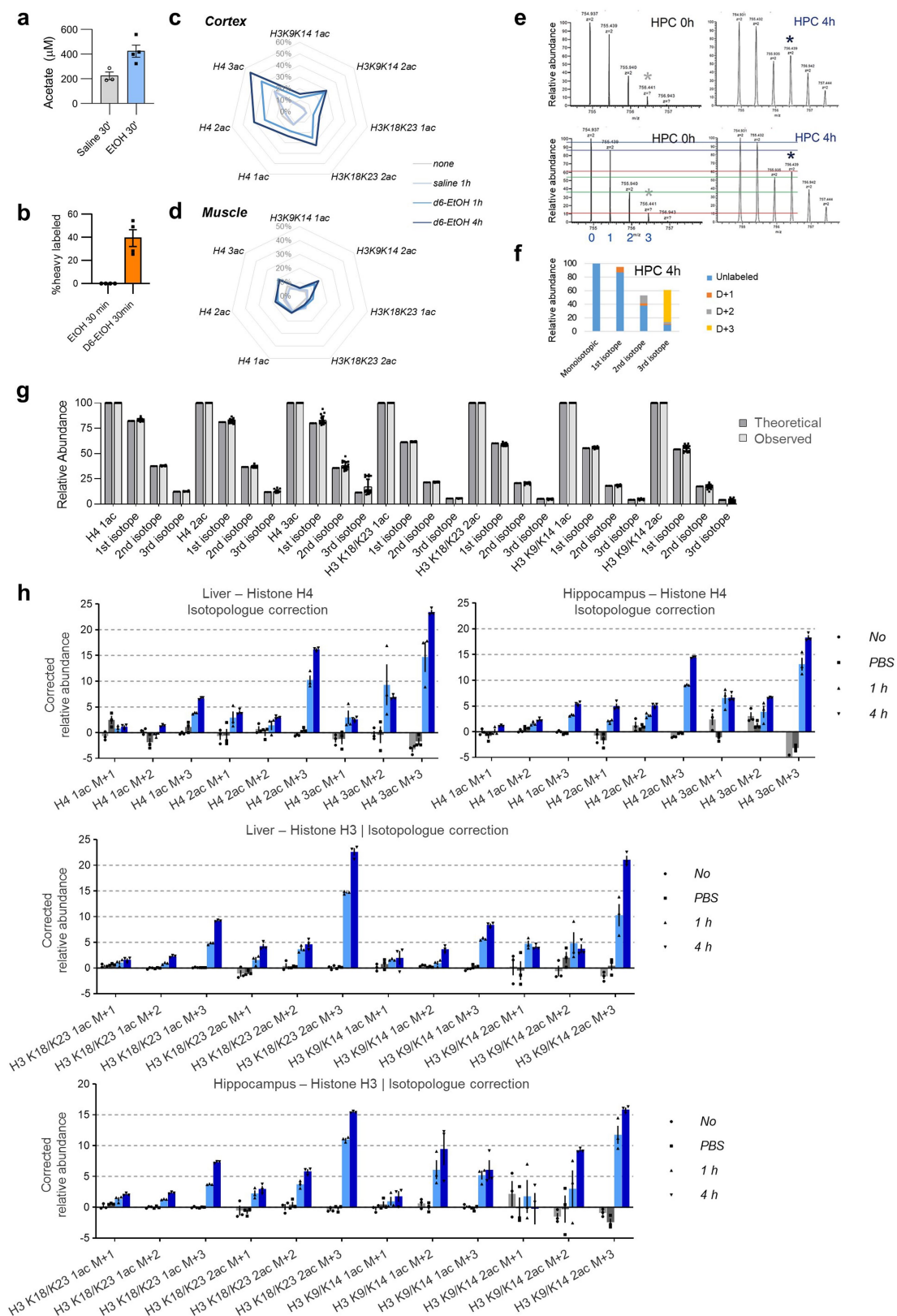
## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1700-7>.

**Correspondence and requests for materials** should be addressed to G.E. or S.L.B.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



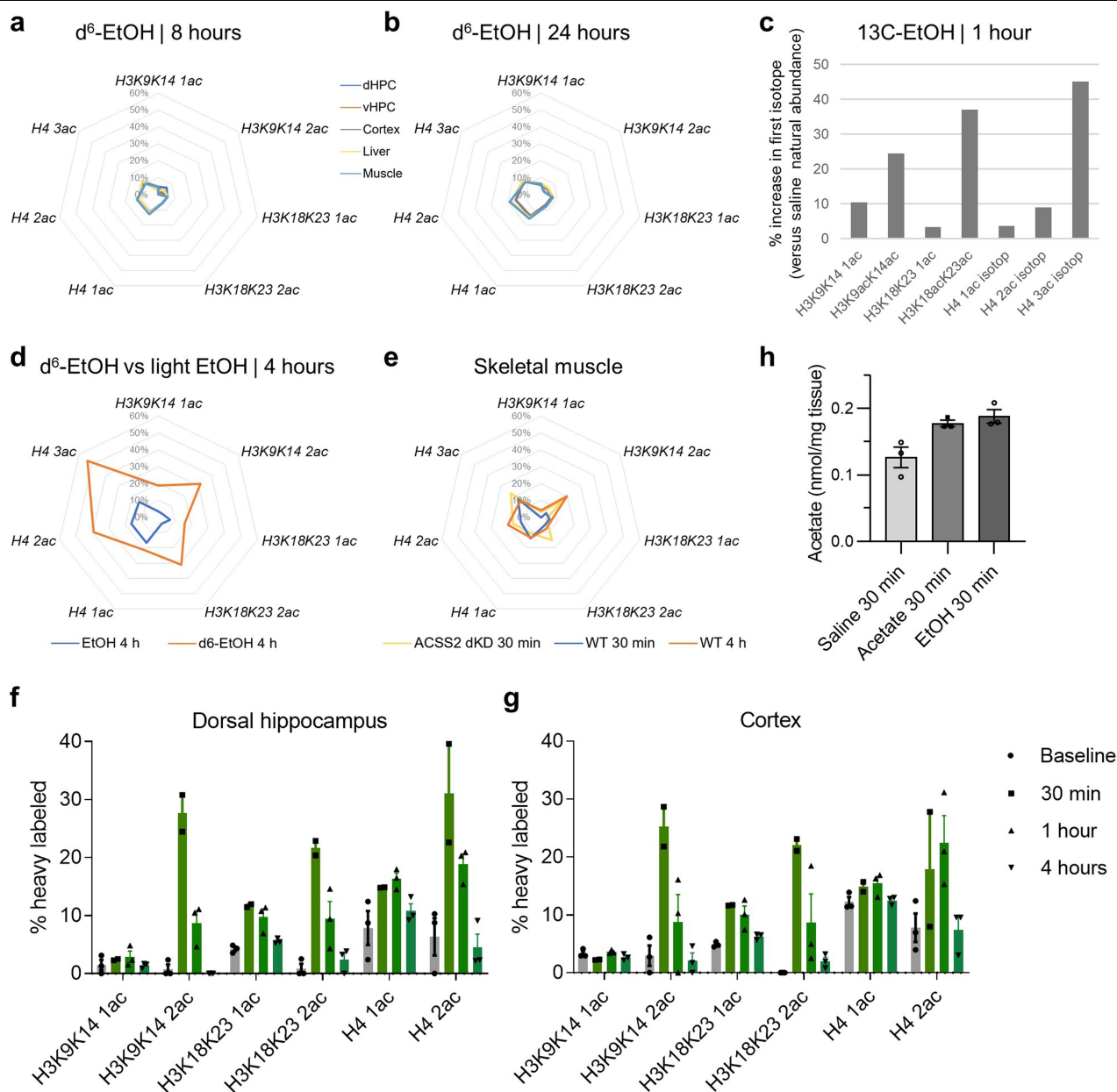


**Extended Data Fig. 1** | See next page for caption.

## Extended Data Fig. 1 | Ethanol-derived acetyl groups are rapidly

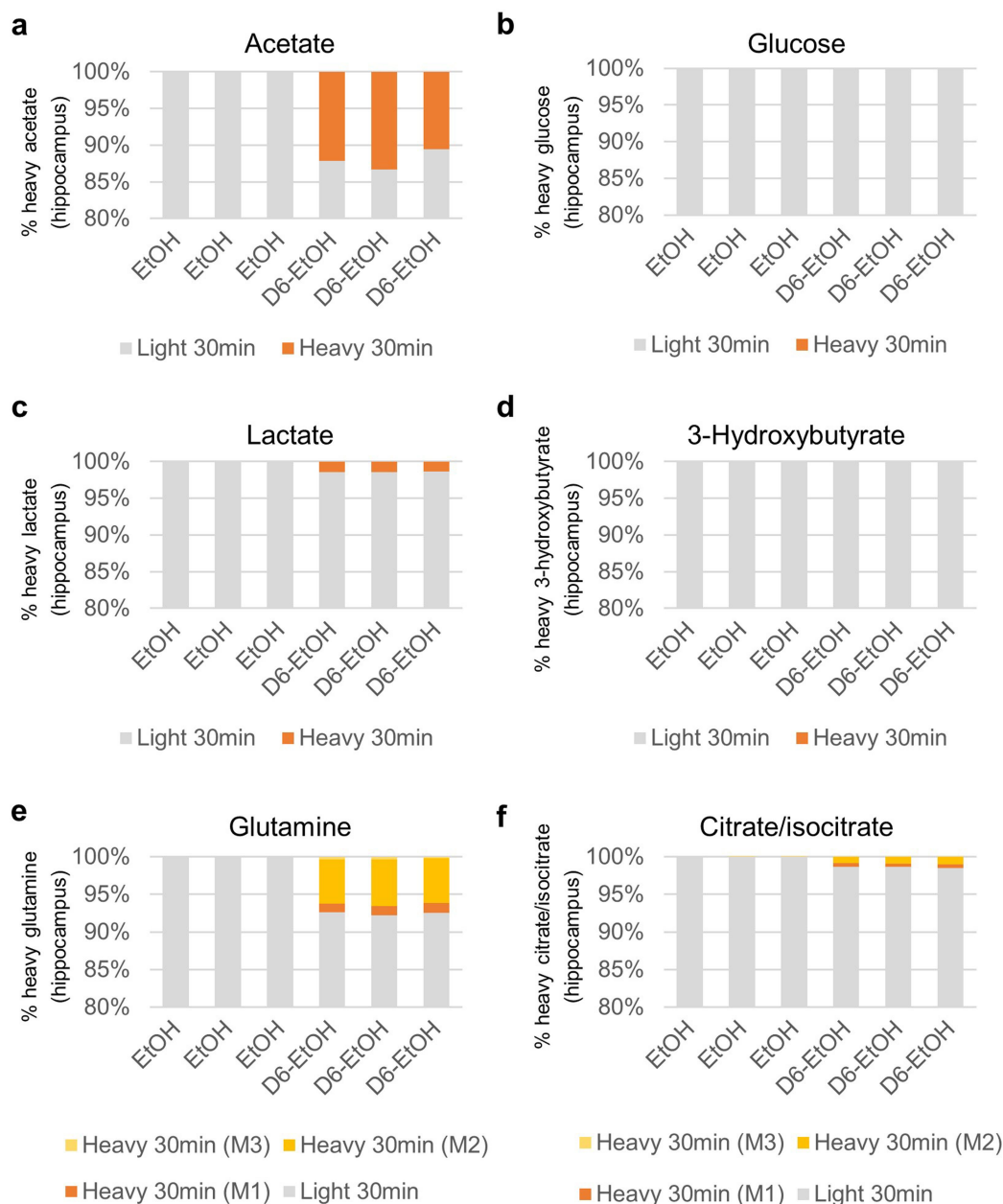
**incorporated into histone acetylation in the brain.** **a**, Mass spectrometry analysis of serum acetate shows the rapid increase in levels of acetate in mice that were injected with alcohol, at 30 min after injection ( $n = 3$  for saline,  $n = 4$  for acetate group). Data are mean  $\pm$  s.e.m.;  $P = 0.0258$  (two-tailed unpaired  $t$ -test). **b**, Ethanol-d6 is readily metabolized and thus labels blood acetate pools. Acetate-d3 was detected by mass spectrometry ( $n = 4$  per group). Data are mean  $\pm$  s.e.m.;  $P = 0.0016$  (two-tailed unpaired  $t$ -test). **c**, Incorporation of the ethanol-d6 label into histone acetylation in the cortex shows a similar pattern to the hippocampus. The axis of the Arachne plot represents the percentage of the third isotope for the acetylated peptide, corresponding to the  $d_3$ -labelled form; the natural relative abundance of that isotope is apparent in the 'none' and 'saline 1 h' treatment groups **d**, Histone acetylation is relatively independent of alcohol metabolism in skeletal muscle, a tissue in which the expression of ACSS2 is low. **e**, **f**, Mass spectra (representative examples from three biological replicates) showing the relative abundance of deuterated histone H4-triacetyl

peptide (amino acids 4–17) in the hippocampus of wild-type mice at baseline and 4 h after injection of ethanol-d6. Increases of the M+1 (blue lines), M+2 (green lines) and M+3 (red lines) species are shown in **e** and indicate a major increase of M+3. The contribution of singly (orange), doubly (grey) and triply (yellow) deuterated peptides to the isotopic distribution is shown in **f**. The relative abundance of the M+3 species is increased by about sixfold at 4 h after injection of ethanol-d6, and is overwhelmingly due to the triply deuterated peptides; by contrast, the contribution of singly and doubly deuterated peptides to the M+3 species is minimal. The experiment was performed with three biological replicates per group. **g**, The relative abundance of the first four isotopes of each of the seven peptides in the untreated samples corresponds to the theoretical isotopic distribution of the peptides (calculated using enviPat<sup>30</sup>; samples not treated with ethanol-d6;  $n = 20$ ). Data are mean  $\pm$  s.d. **h**, Natural abundance-corrected contribution of M+1, M+2 and M+3 species to the labelling of histone acetylation in the liver and hippocampus after intraperitoneal injection of ethanol-d6 (calculated using FluxFix<sup>31</sup>;  $n = 3$  per group). Data are mean  $\pm$  s.d.



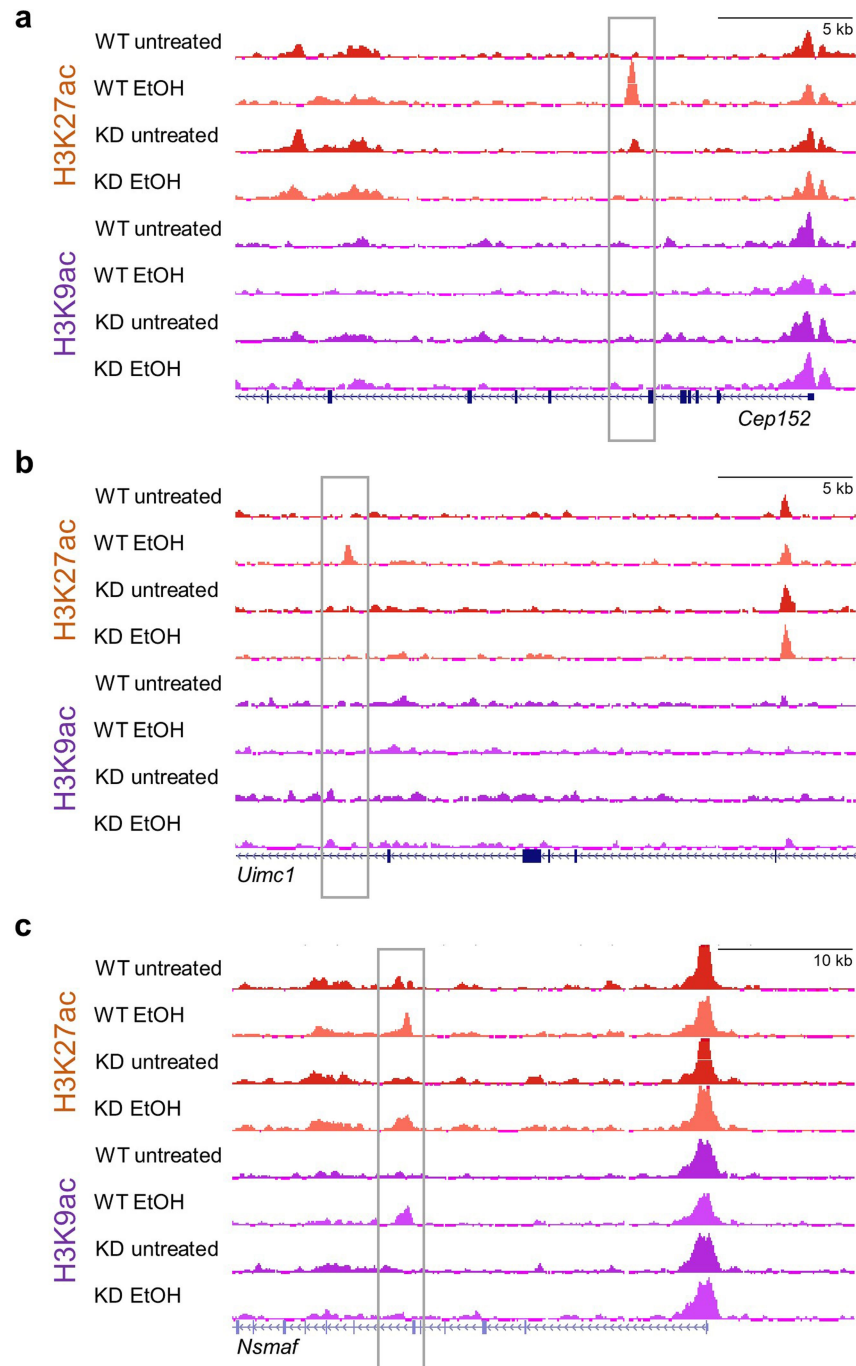
**Extended Data Fig. 2 | Dynamics of ethanol- and acetate-induced heavy-label incorporation.** **a, b**, Relative abundance of deuterated histone acetylation in dHPC, vHPC, cortex, liver and muscle at 8 h (**a**) and 24 h (**b**) after intraperitoneal injection of ethanol-d<sup>6</sup>. **c**, <sup>13</sup>C-labelled ethanol, introduced by intraperitoneal injection, readily labels histone acetylation in the hippocampus (percentage increase over natural abundance of <sup>13</sup>C acetyl groups in saline-injected mice;  $n=1$ ). **d**, In contrast to heavy ethanol-d<sup>6</sup>, a non-labelled ethanol control does not increase the natural abundance of heavy histone acetylation in the hippocampus. **e**, Histone acetylation is relatively independent of alcohol

metabolism in skeletal muscle. Relative abundance of deuterated histone acetylation in skeletal muscle tissue at 30 min and 4 h in wild-type mice, and 30 min in dHPC ACSS2-knockdown mice. **f, g**, Heavy acetate, introduced by intraperitoneal injection, readily labels histone acetylation in the dHPC (**f**) and the cortex (**g**) ( $n=2$  at 30 min;  $n=3$  per group at other time points). Data are mean  $\pm$  s.e.m. **h**, Levels of acetate measured by mass spectrometry in hippocampal tissue after injections of acetate and ethanol ( $n=3$  per group). Data are mean  $\pm$  s.e.m.;  $P=0.0335$  for acetate versus saline;  $P=0.0285$  for ethanol versus saline (two-tailed unpaired  $t$ -test).



**Extended Data Fig. 3 | Metabolite labelling in hippocampal tissue 30 min after intraperitoneal injection of ethanol-d6. a–f,** Mass spectrometry quantification of metabolite labelling in hippocampal tissue at 30 min after intraperitoneal injection of ethanol-d6. The ethanol-d6 label was incorporated

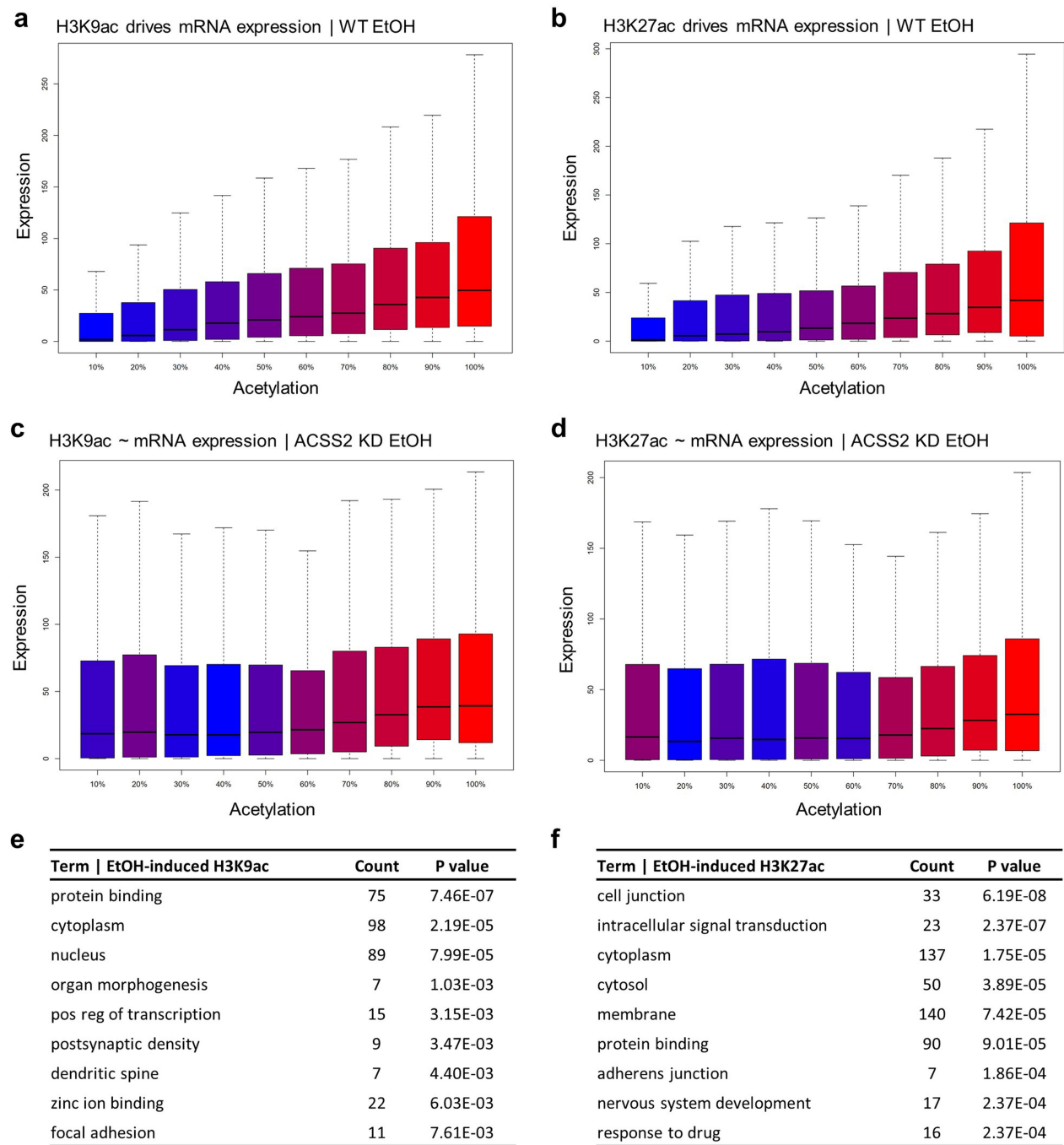
into the acetate pool in the hippocampus (a). By contrast, ethanol-d6 did not contribute to the glucose (b) or 3-hydroxybutyrate (d) pools in the hippocampus, and only minimally to the lactate pool (c). Labelling was observed in the hippocampal glutamine (e) and citrate or isocitrate (f) pools.



**Extended Data Fig. 4 | Representative H3K9ac and H3K27ac dHPC ChIP-seq tracks in control and ethanol-treated wild-type and ACSS2-knockdown mice. a-c,** ChIP-seq for H3K9ac and H3K27ac in untreated and ethanol-treated wild-type and ACSS2-knockdown mice. Genome-browser track views show the

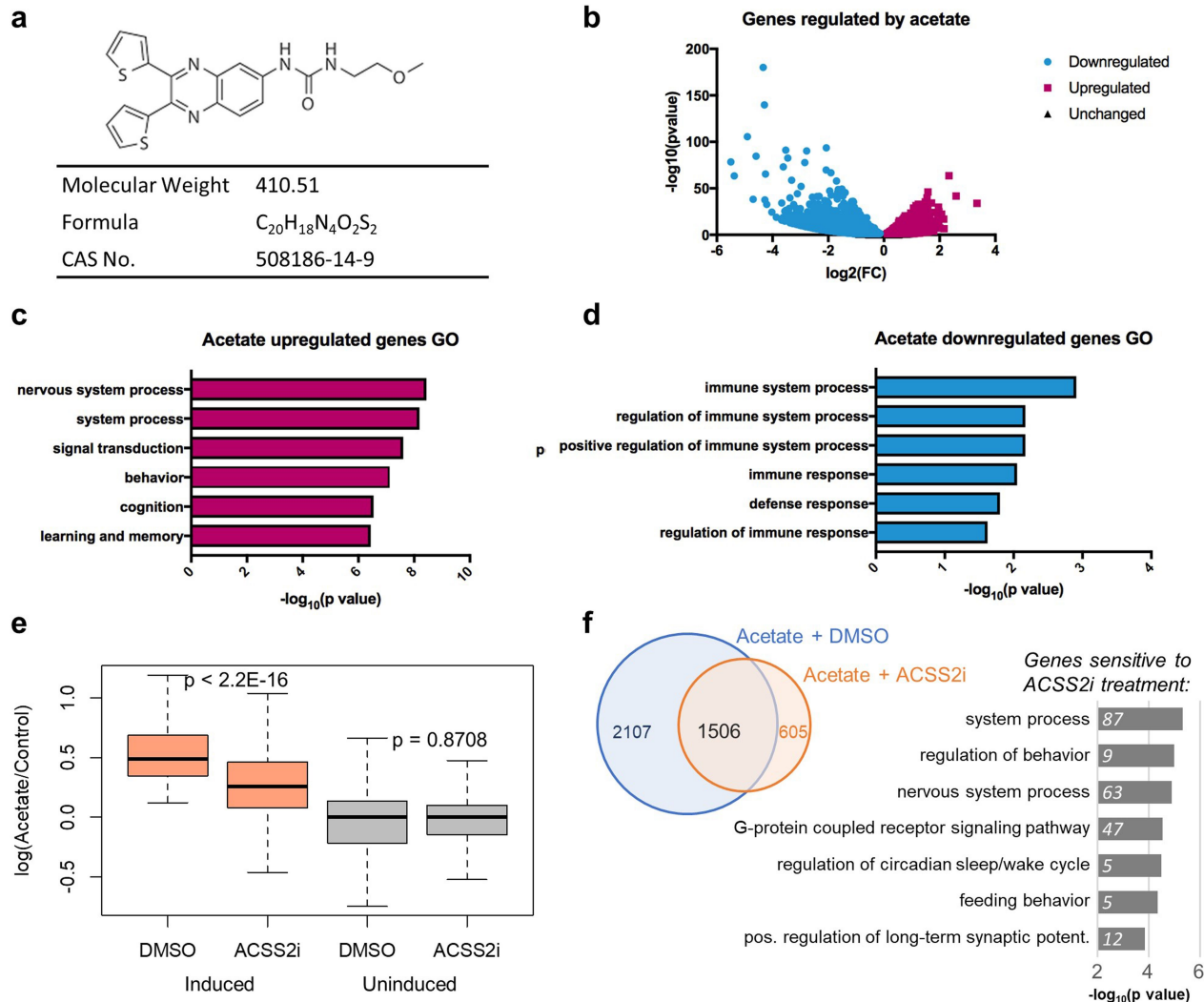
*Cep152* locus (chr2:125,603,000–125,626,000) (a), *Uimc1* locus (chr5:55,064,000–55,089,000) (b) and *Nsmf* locus (chr4:6,425,000–6,464,000) (c). The experiment was performed with three independent biological replicates per group.





**Extended Data Fig. 5 | Epigenetic and transcriptional changes in the dHPC of control and ethanol-treated wild-type and ACSS2-knockdown mice.** **a–d**, Decile plots of genes that are enriched in H3K9ac (**a**) and H3K27ac (**b**) show correlation with mRNA expression levels in hippocampus, in wild-type mice 1 h after injection with ethanol. By contrast, in ACSS2-knockdown mice, the correlation between histone H3K9 acetylation (**c**) and H3K27 acetylation (**d**) and alcohol-related mRNA expression is largely lost ( $n = 16,553$  genes (population)

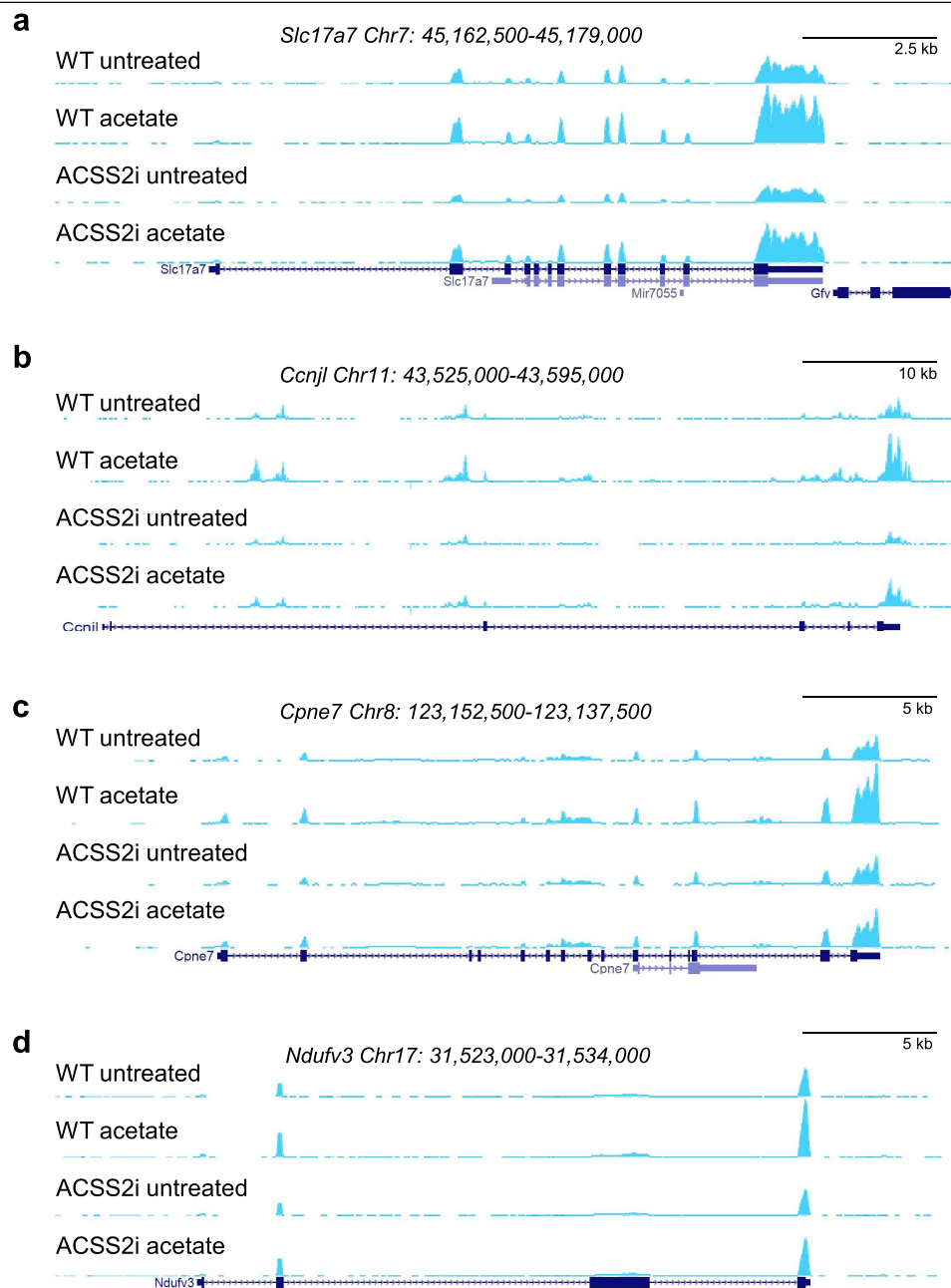
arranged into ten equal-sized deciles by ChIP-seq enrichment of acetylation). Box plots as in Fig. 2. **e, f**, GO analysis on H3K9ac (**e**) and H3K27ac (**f**) peaks that are induced by ethanol in wild-type but not ACSS2-knockdown mice ( $n = 332$  H3K9ac peaks and  $n = 480$  H3K27ac peaks). GO enrichment analysis was performed using a modified Fisher's exact test (EASE) with the FDR corrected by the Yekutieli procedure;  $-\log_{10}$  transformations of nominal  $P$  values are shown.



**Extended Data Fig. 6 | Transcriptional changes in primary hippocampal neurons that were treated with supraphysiological levels of acetate.**

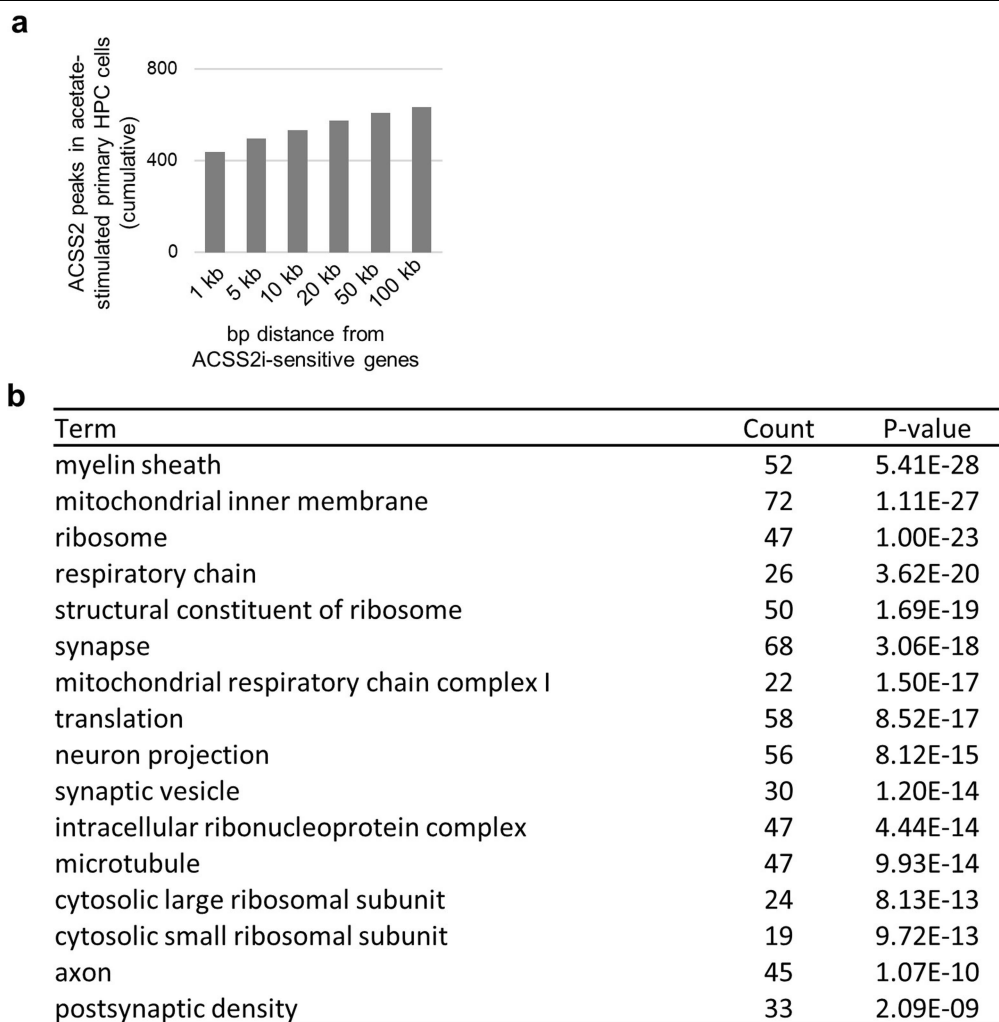
**a**, Structure of ACSS2i ( $C_{20}H_{18}N_4O_2S_2$ ). **b**, RNA-seq showing differentially regulated genes in primary hippocampal neurons that were treated with 5 mM acetate ( $n = 4$  replicates per group; volcano plot of likelihood ratio test (two-sided) in DESeq2; FDR controlled for multiple hypothesis testing). **c**, **d**, GO analysis of genes that are significantly upregulated (**c**;  $n = 3,613$  genes) and significantly downregulated (**d**;  $n = 3,987$  genes) after treatment with 5 mM acetate. GO analysis was performed with GOrilla, using a minimal hypergeometric test. **e**, RNA-seq in primary hippocampal neurons that were isolated from C57/Bl6 mouse embryos and treated with acetate (5 mM) in the

presence or absence of ACSS2i. Of the 3,613 acetate-induced genes, 2,107 are not upregulated in the presence of ACSS2i ( $n = 3,613$  induced genes (population) or 3,613 randomly sampled genes (population);  $P < 2.2 \times 10^{-16}$  (two-sided Mann-Whitney rank-sum test)). Box plots as in Fig. 2. **f**, Acetate-induced genes in primary hippocampal neurons are shown in blue ( $n = 3,613$ ). Of these genes, 2,107 (non-overlapping with orange) were sensitive to ACSS2i, and 1,506 were also induced in the presence of ACSS2i (overlapping with orange). GO enrichment analysis was performed using a modified Fisher's exact test (EASE) with the FDR corrected by the Yekutieli procedure;  $-\log_{10}$  transformations of nominal  $P$  values are shown.



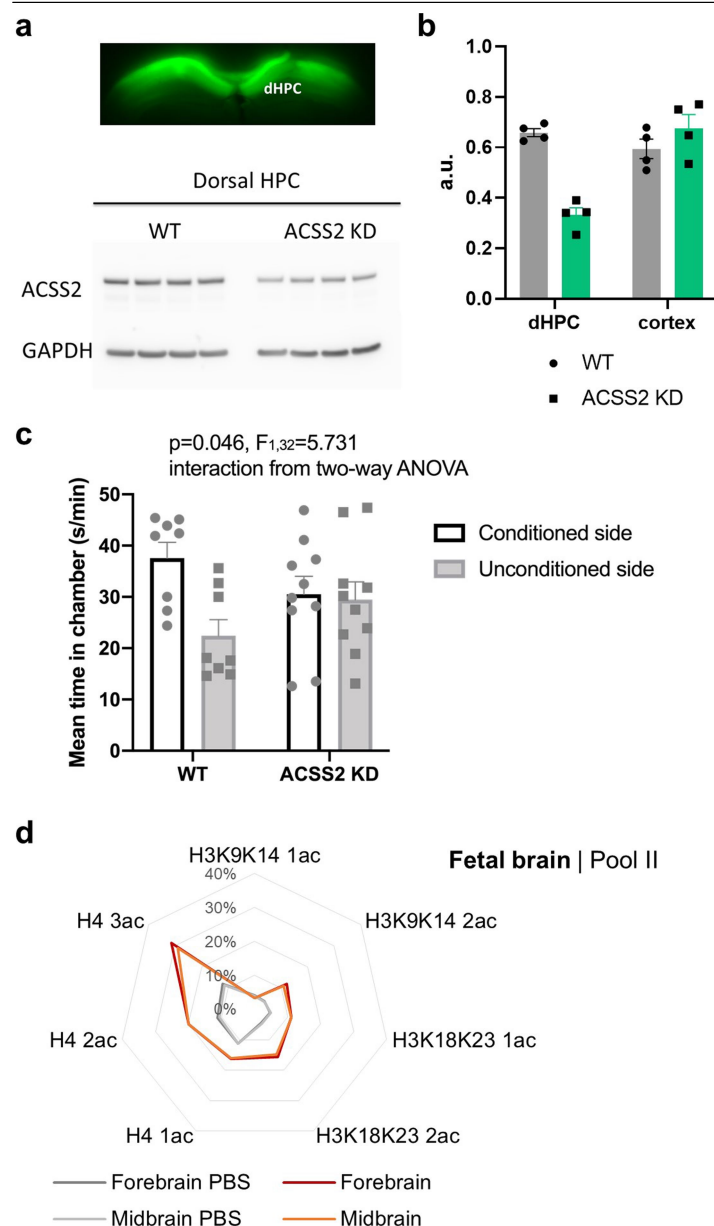
**Extended Data Fig. 7 | Representative RNA-seq tracks in control and acetate-treated primary hippocampal neurons in the presence or absence of ACSS2i. a–d,** Genome-browser track views showing examples of gene upregulation after treatment with acetate in hippocampal neurons, and

diminished induction after treatment with ACSS2i ( $n = 4$  per cohort). RNA-seq track views show the *Slc17a7* locus (chr7: 45,162,500–45,179,000) (a), the *Ccnjl* locus (chr11: 43,525,000–43,595,000) (b), the *Cpne7* locus (chr8: 123,152,500–123,137,500) (c) and the *Ndufv3* locus (chr17: 31,523,000–31,534,000) (d).



**Extended Data Fig. 8 | Transcriptional changes that are induced by acetate in primary hippocampal neurons relate to in vivo ACSS2 peaks and in vivo changes in gene expression that are induced by ethanol. a,** Cumulative number of ACSS2 peaks near the transcription start site of acetylated ACSS2i sensitive genes, indicating that the majority of ACSS2-binding events occur over

or proximal to the gene promoter. **b,** GO analysis for the 830 overlapping genes between the in vivo RNA-seq and ex vivo hippocampal-neuron RNA-seq ( $n = 830$  genes (population)). GO enrichment analysis was performed using a modified Fisher's exact test (EASE) with the FDR corrected by the Yekutieli procedure.



**Extended Data Fig. 9 | Behavioural importance of ACSS2 expression in the dHPC and heavy-label incorporation in the fetal brain.** **a**, Representative image showing virus localization to the dHPC, and western blot ( $n=4$  mice) showing ACSS2 levels in the dHPC of wild-type and ACSS2-knockdown mice (for gel source data, see Supplementary Fig. 1). **b**, Quantification of the levels of ACSS2 protein in the dHPC and cortex of wild-type and dHPC ACSS2-knockdown mice ( $n=4$  mice). Data are mean  $\pm$  s.e.m.;  $P=0.0001$  and  $q=0.0001$  for ACSS2-knockdown versus wild-type mice (dHPC);  $P=0.2666$  and  $q=0.1347$  for ACSS2-knockdown versus wild-type mice (cortex) (multiple  $t$ -test). **c**, ACSS2 is required for alcohol-induced associative learning. Mean time (seconds per minute) spent in unconditioned and ethanol-conditioned chambers following ethanol-induced CPP training in wild-type ( $n=8$ ) and dHPC ACSS2-knockdown mice ( $n=10$ ). Bar graphs represent mean  $\pm$  s.e.m. and data points correspond to individual mice. **d**, Incorporation of ethanol-d6 into histone acetylation in the fetal brain. Data represent the second of two pools of embryos ( $n=4$  per pool) from maternal ethanol-d6 injection. The axes of the Arachne plots represent the percentage of the third isotope for the acetylated peptide, corresponding to the  $d_3$ -labelled form.



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated
- ☐ ☒ Clearly defined error bars  
*State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on [statistics for biologists](#) may be useful.

### Software and code

Policy information about [availability of computer code](#)

Data collection

Software used: STAR v2.5.2a  
HTSeq v0.6.1p1  
DESeq2 v1.22.2  
HOMER v4.6  
DAVID 6.8  
R 3.5.0 for statistical testing  
DiffBind v1.8.5  
MACS2 2.1.0.2010616  
BEDtools v2.15.0  
bedGraphToBigWig v4

Data analysis

GEO SuperSeries GSE122188 [NCBI tracking system #19551484] for detailed information on software used (all publicly available)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing data is accessible on GEO; accession GSE122188

## Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were selected in accordance with the literature and based on previous experience in our group. The Berger lab has experience assessing animal behavior with pharmacological or genetic manipulations and in our experience, robust effects are achieved when group sizes are at least >7-9 animals. Regarding reproducibility, all of our RNA-seq and ChIP-seq datasets were replicated 3-4 times for each condition. Two replicates of RNA-seq are recommended by ENCODE: <a href="https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf">https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf</a> .
Data exclusions	Pre-established exclusion criteria: mice with higher than 65% pre-training to either of the compartments in the CPP were excluded from training. All mice that were trained were included in the analysis (i.e. nothing was excluded after the experiment was run). Grubbs test was used to detect statistically significant outliers. One animal was excluded from the in vivo acetate deposition experiment due to failed i.p. injection (Extended Data Fig 2f/g, 30 min timepoint). No other data were excluded from the analyses.
Replication	RNAseq of primary hippocampal neurons was successfully performed on 4 biological replicates. In vivo alcohol and acetate injection in mice was successfully performed on 2-3 biological replicates per time point. Fetal heavy acetyl group incorporation in mice was successfully tested on 2 biological replicates (two independent pools of fetal brain tissue).
Randomization	All samples and animals were randomly allocated into experimental and control groups.
Blinding	Mass spectrometry analysis of heavy labeled samples was performed by blinded investigators. All animal testing was scored blindly.

## Reporting for specific materials, systems and methods

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

### Methods

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Antibodies

Antibodies used	H3K9ac antibody (Active Motif, cat # 39137; lot # 09811002), H3K27ac antibody (Abcam, cat #4729; lot # GR323132-1)
Validation	ChIP-validated antibodies. Relevant information and citations available on the manufacturers' websites; both antibodies were tested by ChIP analysis. Chromatin IP performed using the ChIP-IT® Express Kit (Catalog No. 53008) and HeLa Chromatin (1.5 x 106 cell equivalents per ChIP) using 10 µl of Histone H3 acetyl Lys9 antibody or the equivalent amount of rabbit IgG as a negative control.

## Animals and other organisms

Policy information about [studies involving animals](#): [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Male and female C57BL/6J mice of 10 weeks of age were used.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.

## ChIP-seq

### Data deposition

- ☒ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links <i>May remain private before publication.</i>	GEO SuperSeries GSE122188 [NCBI tracking system #19551484] Reviewers can access all information using the following secure token: ynupekygbnqnbkl
Files in database submission	all raw sequencing data and processed data discussed in manuscript
Genome browser session (e.g. <a href="#">UCSC</a> )	link in GEO accession

### Methodology

Replicates	n = 4 per condition; 16 total 16 total
Sequencing depth	Libraries were multiplexed using NEBNext Multiplex Oligos for Illumina (dual index primers) and single-ended sequenced (75 bp) on the NextSeq 500 platform (Illumina) in accordance with the manufacturer's protocol. ChIP-seq tags generated with the NextSeq 500 platform were demultiplexed with the bcl2fastq utility and aligned to the mouse reference genome (assembly GRCh38/mm10) using Bowtie v1.1.1 [3] allowing up to two mismatches per sequencing tag (parameters -m 1 --best).
Antibodies	H3K9ac antibody (Active Motif, cat # 39137; lot # 09811002); H3K27ac antibody (Abcam, cat #4729; lot # GR323132-1)
Peak calling parameters	ChIP-seq tags generated with the NextSeq 500 platform were demultiplexed with the bcl2fastq utility and aligned to the mouse reference genome (assembly GRCh38/mm10) using Bowtie v1.1.1 [3] allowing up to two mismatches per sequencing tag (parameters -m 1 --best). Peaks were detected using MACS2 (tag size = 75 bp; FDR < 1x10 <sup>-3</sup> ) from pooled H3K9ac or H3K27ac tags of mice from the same condition along with treatment-matched Input tags as control.
Data quality	The MTL method was used to compare H3K9ac or H3K27ac enrichment in the four study conditions. Statistical significance of differential H3K9ac or H3K27ac enrichments was assessed by using DiffBind (Bioconductor v3.7) in a 2-way comparison (wt-saline vs wt-EtOH or ACSS2KD-saline vs ACSS2KD-EtOH) across the individual replicate samples (FDR < 10%).
Software	UCSC Genome Browser track views were created for ChIP-seq data by first pooling replicates and generating coverage maps using BEDtools genomeCoverageBed -bg, then adjusting for library size using the RPM coefficient. Input signal was then subtracted from ChIP signal. Resulting tracks were converted from bedGraph to bigWig using the Genome Browser's bedGraphToBigWig utility. RNA-seq tracks were created similarly, first splitting by tag orientation to the genomic reference strand and then creating coverage maps. Because an RPM adjustment might disguise a large deformation in the transcriptome distribution, maps were adjusted for library size using the average scalar coefficient size factor determined by DESeq2. Resulting tracks were converted to bigWigs as ChIP-seq tracks were, and the + and - tags from a given sample were plotted as overlays in a track hub.

# Structural basis for enzymatic photocatalysis in chlorophyll biosynthesis

<https://doi.org/10.1038/s41586-019-1685-2>

Received: 22 May 2019

Accepted: 25 September 2019

Published online: 23 October 2019

Shaowei Zhang<sup>1,7</sup>, Derren J. Heyes<sup>1,7</sup>, Lingling Feng<sup>2,7</sup>, Wenli Sun<sup>3,7</sup>, Linus O. Johannissen<sup>1,7</sup>, Huanling Liu<sup>4</sup>, Colin W. Levy<sup>1</sup>, Xuemei Li<sup>5</sup>, Ji Yang<sup>4</sup>, Xiaolan Yu<sup>4</sup>, Min Lin<sup>3</sup>, Samantha J. O. Hardman<sup>1</sup>, Robin Hoeven<sup>1</sup>, Michiyo Sakuma<sup>1</sup>, Sam Hay<sup>1</sup>, David Leys<sup>1</sup>, Zihe Rao<sup>5</sup>, Aiwu Zhou<sup>2\*</sup>, Qi Cheng<sup>3,4,6\*</sup> & Nigel S. Scrutton<sup>1\*</sup>

The enzyme protochlorophyllide oxidoreductase (POR) catalyses a light-dependent step in chlorophyll biosynthesis that is essential to photosynthesis and, ultimately, all life on Earth<sup>1–3</sup>. POR, which is one of three known light-dependent enzymes<sup>4,5</sup>, catalyses reduction of the photosensitizer and substrate protochlorophyllide to form the pigment chlorophyllide. Despite its biological importance, the structural basis for POR photocatalysis has remained unknown. Here we report crystal structures of cyanobacterial PORs from *Thermosynechococcus elongatus* and *Synechocystis* sp. in their free forms, and in complex with the nicotinamide coenzyme. Our structural models and simulations of the ternary protochlorophyllide–NADPH–POR complex identify multiple interactions in the POR active site that are important for protochlorophyllide binding, photosensitization and photochemical conversion to chlorophyllide. We demonstrate the importance of active-site architecture and protochlorophyllide structure in driving POR photochemistry in experiments using POR variants and protochlorophyllide analogues. These studies reveal how the POR active site facilitates light-driven reduction of protochlorophyllide by localized hydride transfer from NADPH and long-range proton transfer along structurally defined proton-transfer pathways.

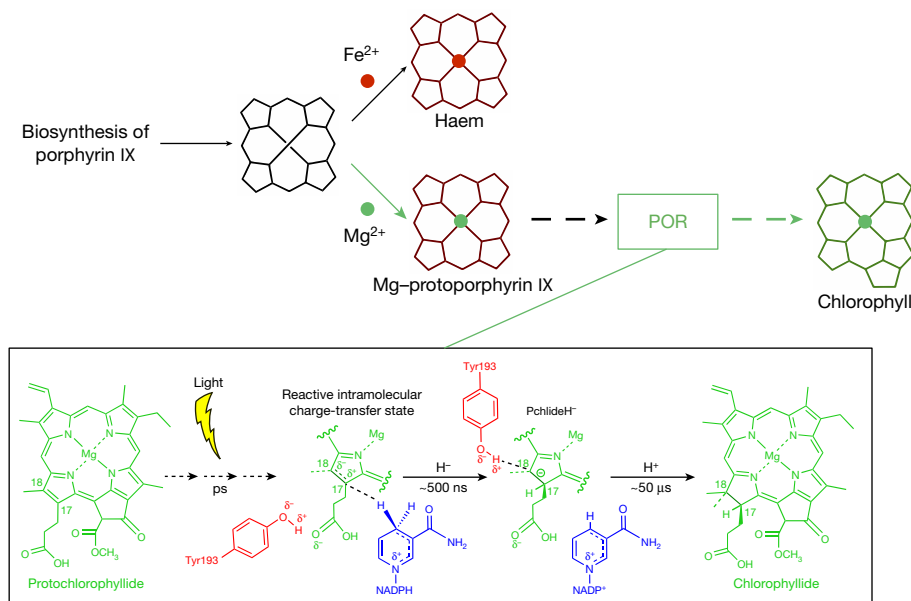
As the light-driven step in the chlorophyll biosynthetic pathway (Fig. 1), the POR reaction acts as the trigger for the germination of seedlings in plants and provokes a marked change in the morphological development of the plant<sup>2,3</sup>. Given this crucial biological role, POR has been the focus of numerous mechanistic and biophysical investigations. A combination of time-resolved (at the femtosecond-to-second scale) and cryogenic spectroscopy methods have provided some understanding of the mechanism of POR photocatalysis in a range of photosynthetic organisms, including cyanobacteria and plants. Picosecond excited-state dynamics in the protochlorophyllide (Pchlde) molecule are thought to result in excited-state interactions between the substrate and active-site residues that are necessary to trigger the subsequent reaction chemistry<sup>6–12</sup>. This involves the sequential transfer of a hydride equivalent from NADPH and a proton transfer from either an active-site residue or solvent. Proton transfer is reliant on solvent dynamics and an implied network of extended protein motions that occur on the microsecond timescale<sup>13–17</sup>. Hydride transfer from NADPH is not concerted but instead occurs in a stepwise manner that involves Pchlde excited-state electron transfer from NADPH followed by a proton-coupled electron transfer<sup>18</sup>, which represents the first example of the stepwise transfer of a hydride equivalent that has been reported in biology. These time-resolved studies have provided insight into the chemistry of catalysis across a wide range of timescales

(from the femtosecond to the second), but the required structural basis of POR photocatalysis has remained unknown. This structural context is required to understand how substrate binding, excited-state chemistry, bond making and/or breaking and the dynamics of photocatalysis are controlled by protein structure.

Here we report the crystal structures of the apo-POR enzyme from *T. elongatus* (RCSB Protein Data Bank (PDB) code 6RNV) and NADPH-bound POR from both *Synechocystis* sp. and *T. elongatus* (PDB codes 6G08 and 6RNW, respectively) (Extended Data Table 1a, Supplementary Methods), solved at 1.3 Å, 1.9 Å and 1.9 Å resolution, respectively (Fig. 2, Extended Data Fig. 1). As a member of the short-chain dehydrogenase and reductase (SDR) family of enzymes, the overall structure of POR is similar to that of other members of the SDR family<sup>19</sup> and has a typical dinucleotide-binding Rossmann fold, which comprises a central  $\beta$ -sheet surrounded by six  $\alpha$ -helices (Extended Data Figs. 1, 2). Three flexible regions of the structure (residues 146–160, 228–255 and 284–291; hereafter denoted R1, R2 and R3, respectively) are not observed in the *T. elongatus* electron density. This observation is consistent with molecular-dynamics simulations of both the apo- and coenzyme-bound structures (Extended Data Fig. 3), which indicates that these regions have a high degree of flexibility. Despite this, both the R1 region and the R2 region are ordered in the coenzyme-bound structure of the *Synechocystis*

<sup>1</sup>Manchester Institute of Biotechnology, The University of Manchester, Manchester, UK. <sup>2</sup>Department of Pathophysiology, Key Laboratory of Cell Differentiation and Apoptosis of the Chinese Ministry of Education, Shanghai Jiao Tong University School of Medicine, Shanghai, China. <sup>3</sup>Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing, China.

<sup>4</sup>C4-101, Nitrogen Fixation Laboratory, Qi Institute, Jiaying, China. <sup>5</sup>National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. <sup>6</sup>College of Biotechnology and Bioengineering, Zhejiang University of Technology, Hangzhou, China. <sup>7</sup>These authors contributed equally: Shaowei Zhang, Derren J. Heyes, Lingling Feng, Wenli Sun, Linus O. Johannissen. \*e-mail: awz20@shsmu.edu.cn; chengqi@caas.cn; nigel.scrutton@manchester.ac.uk



**Fig. 1 | The light-driven reduction of the C-17–C-18 double bond of Pchlide to form chlorophyllide.** The reaction is catalysed by POR and is a key regulatory step within the chlorophyll biosynthetic pathway. Catalysis involves excited-

state interactions between the Pchlide and the protein, which lead to sequential hydride transfer from NADPH to the C-17 position and proton transfer to the C-18 position.

POR and therefore implicated in coenzyme binding to POR. A short loop (residues 223–229) within the R2 region extends from the central  $\beta$ -sheet, covers the nicotinamide moiety of the NADPH and becomes ordered upon binding of the coenzyme. A similar loop is observed over the coenzyme-binding pocket of other SDR enzymes, and it has previously been shown that coenzyme binding causes the extended area around the loop to form two short  $\alpha$ -helices that act as a lid to cover the active site, which leads to a ‘closed’ conformation<sup>20–23</sup>. In this region of the *Synechocystis* POR, only one longer  $\alpha$ -helix (residues 230–239; hereafter referred to as helix 1) (Fig. 2a) is observed—although an additional loop (residues 148–159; hereafter referred to as loop 1) (Fig. 2a), which extends from the central  $\beta$ -sheet, is located near helix 1. It appears that loop 1, which is conserved in POR enzymes (Extended Data Fig. 1e, Supplementary Fig. 1), and helix 1 are important in controlling the subsequent binding of the Pchlide substrate.

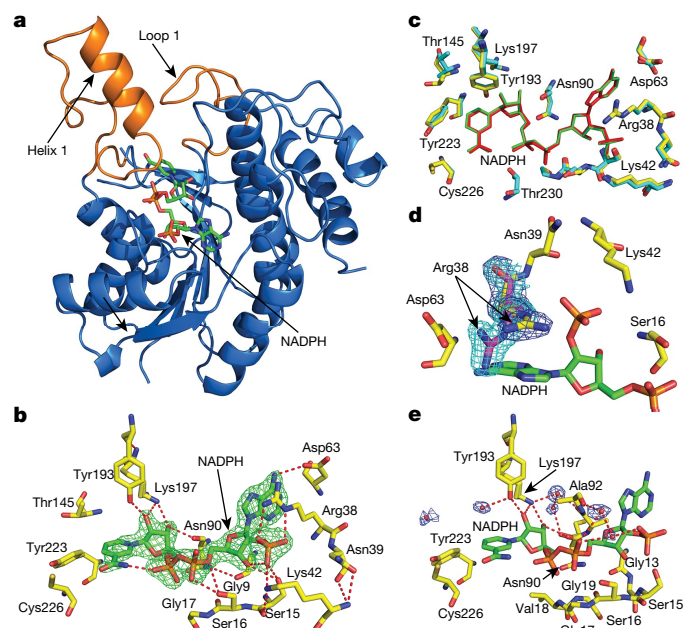
Beyond the loop regions, our POR crystal structures identify additional residues that are important to coenzyme binding. The SDR family of proteins uses an Asn-Ser-Tyr-Lys catalytic tetrad for proton transfer and stabilization of reaction intermediates<sup>19</sup>; however, in POR a Thr residue (Thr145) often replaces the Ser residue (Supplementary Fig. 1). Crystal structures of the binary POR–NADPH complex show that three of these residues—Asn90, Tyr193 and Lys197 (numbering refers to the *T. elongatus* POR)—are directly hydrogen-bonded to the coenzyme (Fig. 2b, c). Also, Arg38, Lys42 and Asp63 interact with the coenzyme via a hydrogen-bonding network to the 2′ phosphate of the molecule (Fig. 2b, c, Extended Data Fig. 4). Arg38 is implicated in coenzyme binding and release as it adopts alternative conformations in the different POR structures (Fig. 2d), a hypothesis that is further supported by the reduced NADPH binding affinity observed in an R38V variant<sup>24</sup>. A water channel surrounding the Tyr193 and Lys197 residues of the active site is also identified in our structures, which could act as a potential proton relay network during catalysis (Fig. 2e).

Despite extensive efforts, crystals of the ternary POR complex (comprising POR, Pchlide and NADP(H)) were not obtained, as a result of the enzyme adopting an array of oligomeric states on binding the substrate (Supplementary Fig. 2). We therefore used molecular docking to produce an initial model for the structure of the POR–Pchlide–NADPH ternary complex, followed by simulated annealing molecular-dynamics simulations and umbrella sampling to calculate the potential of mean force

for Pchlide binding (Extended Data Figs. 3, 5). Although the potential of mean force changes over the time course of the simulation, the lowest energy conformation and the approximate dissociation energy remain stable and well-defined (Extended Data Fig. 5d, e). The resulting binding energy of about 35 kJ mol<sup>−1</sup> (which corresponds to a dissociation constant of 0.8  $\mu$ M) agrees well with the experimental dissociation constant of about 2  $\mu$ M (Extended Data Table 1b, Supplementary Fig. 3). The lowest energy conformation yields a ternary complex structure that is more rigid than the flexible, highly mobile structures that were observed for apo-POR and POR–NADPH (Extended Data Fig. 3a–c). In the ternary complex structure, helix 1 folds over the Pchlide and, together with loop 1, forms a lid over the active site (Fig. 3a). The hydrophobic edge of Pchlide points towards hydrophobic regions of the enzyme (including the residues in helix 1), and the keto, methylester and carboxylic acid groups of the Pchlide molecule are positioned in a hydrophilic pocket (Fig. 3b). This provides numerous hydrogen-bonding opportunities and positions Pchlide in an orientation that is consistent with the required stereochemistry of hydride and proton transfer (Fig. 3c, Extended Data Fig. 5). Specific interactions include a salt bridge between Lys197 and the carboxylic acid side chain at the C-17 position, which also forms a hydrogen bond to Thr145. The central Mg<sup>2+</sup> ion of Pchlide is ligated to two water molecules, one of which appears to be part of a water network that also involves hydrogen bonds to Tyr223 and the methylester at the C-15 position.

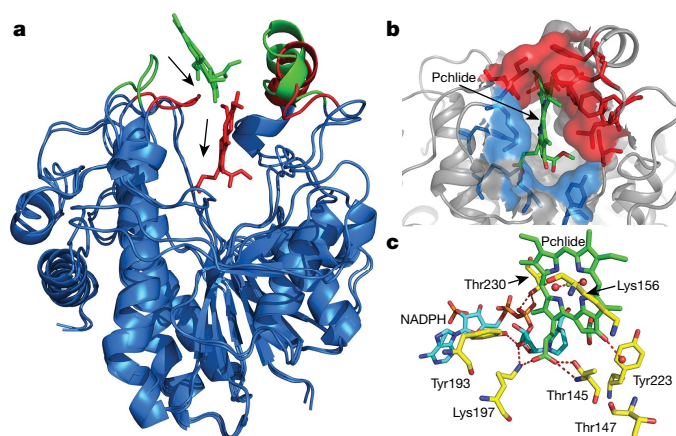
The photocatalytic implications of the determined structures and models of the active ternary complex were investigated using activity, binding and inhibition studies of site-directed variants and analogues of Pchlide with modifications to substituent groups that are of potential importance to the enzyme mechanism (Extended Data Fig. 6, Extended Data Table 1b, Supplementary Figs. 3–8). Pchlide analogues were altered in the nature of the substituents on ring E, the central Mg<sup>2+</sup> ion and the carboxylic acid side chain at the C-17 position, all of which have previously been inferred as being important in POR-catalysed photoreduction<sup>25</sup>. An analogue in which the propionic acid side chain at the C-17 position is exchanged for a methylester group does not bind to POR. This analogue is also not a competitive inhibitor of the natural enzyme-catalysed reaction. This emphasizes the importance of this propionic acid side chain in Pchlide binding to the POR active site, consistent with our structural models. Mutagenesis of Lys197 or





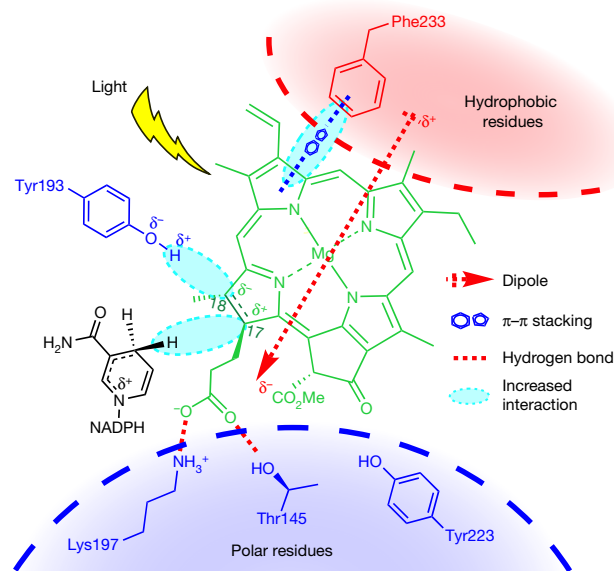
Thr145 leads to an impaired ability of POR to bind Pchlide substrate<sup>24,26</sup>. This finding is consistent with the identified salt bridge and hydrogen bond made by these residues in their interaction with the propionic acid side chain of Pchlide. Additional Pchlide analogues modified at a range of locations of the porphyrin retain an ability to bind to the enzyme. These analogues also act as competitive inhibitors, albeit with affinities that are reduced compared to Pchlide itself (Extended Data Fig. 6, Extended Data Table 1b). Loss of the central  $Mg^{2+}$  ion and the keto group (C-13 position) led to about a 3–5 fold increase in  $K_d$ , as expected from proposed roles in binding via coordination with water molecules in the active site. Although Lys156 is situated in the loop region of the lid (in close proximity to Pchlide), activity measurements using a K156A variant indicate that Lys156 is not important for binding (Extended Data Table 1b, Supplementary Fig. 8). This is consistent with the lack of hydrogen-bonding from Lys156 to Pchlide in the final model of the ternary complex. Molecular-dynamics simulations indicate Tyr223 facilitates the binding of Pchlide in the orientation that is required for catalysis via hydrogen bonding to the keto group at an intermediary phase in the binding process ( $R_0$  of about 17 Å in the potential of mean force in Extended Data Figs. 5, 7) but does not directly interact with Pchlide in the ternary complex (Extended Data Fig. 6, Supplementary Video 1). This is consistent with experimental binding data, in which a considerable reduction of about four- and tenfold is observed in the binding of Pchlide to Y223F and Y223A variants of the PORs, respectively (Extended Data Table 1b, Supplementary Fig. 8).

Thr145 leads to an impaired ability of POR to bind Pchlide substrate<sup>24,26</sup>. This finding is consistent with the identified salt bridge and hydrogen bond made by these residues in their interaction with the propionic acid side chain of Pchlide. Additional Pchlide analogues modified at a range of locations of the porphyrin retain an ability to bind to the enzyme. These analogues also act as competitive inhibitors, albeit with affinities that are reduced compared to Pchlide itself (Extended Data Fig. 6, Extended Data Table 1b). Loss of the central  $Mg^{2+}$  ion and the keto group (C-13 position) led to about a 3–5 fold increase in  $K_d$ , as expected from proposed roles in binding via coordination with water molecules in the active site. Although Lys156 is situated in the loop region of the lid (in close proximity to Pchlide), activity measurements using a K156A variant indicate that Lys156 is not important for binding (Extended Data Table 1b, Supplementary Fig. 8). This is consistent with the lack of hydrogen-bonding from Lys156 to Pchlide in the final model of the ternary complex. Molecular-dynamics simulations indicate Tyr223 facilitates the binding of Pchlide in the orientation that is required for catalysis via hydrogen bonding to the keto group at an intermediary phase in the binding process ( $R_0$  of about 17 Å in the potential of mean force in Extended Data Figs. 5, 7) but does not directly interact with Pchlide in the ternary complex (Extended Data Fig. 6, Supplementary Video 1). This is consistent with experimental binding data, in which a considerable reduction of about four- and tenfold is observed in the binding of Pchlide to Y223F and Y223A variants of the PORs, respectively (Extended Data Table 1b, Supplementary Fig. 8).



The POR crystal structures and ternary complex model provide a structural basis for understanding light-activated catalysis in POR. The extensive hydrogen-bonding network identified between active-site residues and Pchlide is likely to be essential for photochemistry, and these interactions are proposed to strengthen in the excited state to create an electron-deficient site at the C-17–C-18 double bond<sup>9,10</sup>. The central  $Mg^{2+}$  ion and keto group at the C-13 position of the Pchlide are also important in Pchlide photochemistry (Extended Data Fig. 8, Supplementary Fig. 9), as they facilitate charge separation across the Pchlide during photoexcitation<sup>11</sup>. The active-site architecture of POR (Extended Data Fig. 9) is finely tuned to facilitate this excited-state charge separation and to stabilize the strengthened dipole across the Pchlide molecule (Fig. 4). The positive end of the dipole is located within the hydrophobic pocket of the active site,

The POR crystal structures and ternary complex model provide a structural basis for understanding light-activated catalysis in POR. The extensive hydrogen-bonding network identified between active-site residues and Pchlide is likely to be essential for photochemistry, and these interactions are proposed to strengthen in the excited state to create an electron-deficient site at the C-17–C-18 double bond<sup>9,10</sup>. The central  $Mg^{2+}$  ion and keto group at the C-13 position of the Pchlide are also important in Pchlide photochemistry (Extended Data Fig. 8, Supplementary Fig. 9), as they facilitate charge separation across the Pchlide during photoexcitation<sup>11</sup>. The active-site architecture of POR (Extended Data Fig. 9) is finely tuned to facilitate this excited-state charge separation and to stabilize the strengthened dipole across the Pchlide molecule (Fig. 4). The positive end of the dipole is located within the hydrophobic pocket of the active site,



**Fig. 4 | A structural basis for light-dependent reaction chemistry in POR.** Mechanistic scheme that illustrates the potential structural reorganization of the Pchlide-binding site upon excitation.

in which it is stabilized by  $\pi$ - $\pi$  stacking interactions with a conserved Phe residue. By contrast, the negative end of the dipole is found in a region of polar residues. Interactions between Thr145 and Lys197 of POR and the propionic acid side chain of Pchl<sub>id</sub> at C-17 are important to excited-state chemistry, as shown by the fact that changes to either residue results in impaired photochemistry<sup>10,24,26</sup>. These excited-state interactions between Pchl<sub>id</sub> and POR stabilize a highly polarized C-17–C-18 double bond and enable stepwise hydride transfer from NADPH to C-17<sup>18</sup>. The polarized nature of the C-17–C-18 bond may also be stabilized by the close proximity of the hydroxyl group of Tyr193, which is known to be required for Pchl<sub>id</sub> photochemistry<sup>10</sup>. By contrast, Tyr223 does not form any direct interactions with Pchl<sub>id</sub> (Extended Data Fig. 6, Extended Data Table 1b) and—consistent with this finding—the POR variants Y223A and Y223F are able to catalyse hydride transfer from NADPH to Pchl<sub>id</sub> (Supplementary Figs. 10, 11). The donor–acceptor distance between the NADPH *pro*-S hydrogen and Pchl<sub>id</sub> C-17 is  $4.5 \pm 0.3$  Å (Extended Data Fig. 6b), but this may change slightly upon photoexcitation of the Pchl<sub>id</sub> molecule. The subsequent transfer of a second proton to the C-18 of Pchl<sub>id</sub> is from the strictly conserved Tyr193 of POR<sup>14,26</sup> (Extended Data Fig. 6c) with a donor–acceptor distance of  $4.9 \pm 0.4$  Å in our model—although this distance may change upon formation of a Pchl<sub>id</sub> anion following reduction of the C-17–C-18 double bond. In reactions catalysed by other members of the SDR family, the catalytic Tyr is replenished through a proton relay mechanism<sup>19</sup>. On the basis of the crystal structures reported here, the adjacent Lys197 is implicated in modulating the ionization properties of Tyr193. As the Y223A and Y223F variants of POR possess reduced rates of proton transfer (Supplementary Fig. 11), the water network coordinated by Tyr223 may also be important in this proton relay mechanism. The crystal structure of POR reported here should enable computational and time-resolved structural–mechanistic studies of the complete enzyme reaction cycle to provide spatial, temporal and energetic understanding across multiple timescales (for example, femtosecond-to-second), and thereby address a major challenge in biological catalysis.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

The atomic coordinates and experimental data have been deposited in the Protein Data Bank with accession codes 6G08, 6RNV and 6RNW. All other data are available from the corresponding authors on reasonable request.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1685-2>.

1. Yang, J. & Cheng, Q. Origin and evolution of the light-dependent protochlorophyllide oxidoreductase (LPOR) genes. *Plant Biol.* **6**, 537–544 (2004).
2. Scrutton, N. S., Groot, M. L. & Heyes, D. J. Excited state dynamics and catalytic mechanism of the light-driven enzyme protochlorophyllide oxidoreductase. *Phys. Chem. Chem. Phys.* **14**, 8818–8824 (2012).

3. Gabruk, M. & Mysliwa-Kurczel, B. Light-dependent protochlorophyllide oxidoreductase: phylogeny, regulation, and catalytic properties. *Biochemistry* **54**, 5255–5262 (2015).
4. Aubert, C., Vos, M. H., Mathis, P., Eker, A. P. & Brettel, K. Intraprotein radical transfer during photoactivation of DNA photolyase. *Nature* **405**, 586–590 (2000).
5. Sorigué, D. et al. An algal photoenzyme converts fatty acids to hydrocarbons. *Science* **357**, 903–907 (2017).
6. Dietzek, B. et al. Excited-state processes in protochlorophyllide a: a femtosecond time-resolved absorption study. *Chem. Phys. Lett.* **397**, 110–115 (2004).
7. Dietzek, B. et al. Dynamics of charge separation in the excited-state chemistry of protochlorophyllide. *Chem. Phys. Lett.* **492**, 157–163 (2010).
8. Sytina, O. A. et al. Protochlorophyllide excited-state dynamics in organic solvents studied by time-resolved visible and mid-infrared spectroscopy. *J. Phys. Chem. B* **114**, 4335–4344 (2010).
9. Zhao, G. J. & Han, K. L. Site-specific solvation of the photoexcited protochlorophyllide in methanol: formation of the hydrogen-bonded intermediate state induced by hydrogen-bond strengthening. *Biophys. J.* **94**, 38–46 (2008).
10. Heyes, D. J. et al. Excited-state charge separation in the photochemical mechanism of the light-driven enzyme protochlorophyllide oxidoreductase. *Angew. Chem. Int. Ed.* **54**, 1512–1515 (2015).
11. Heyes, D. J. et al. Excited-state properties of protochlorophyllide analogues and implications for light-driven synthesis of chlorophyll. *J. Phys. Chem. B* **121**, 1312–1320 (2017).
12. Brandariz-de-Pedro, G. et al. Direct evidence of an excited-state triplet species upon photoactivation of the chlorophyll precursor protochlorophyllide. *J. Phys. Chem. Lett.* **8**, 1219–1223 (2017).
13. Heyes, D. J. et al. The first catalytic step of the light-driven enzyme protochlorophyllide oxidoreductase proceeds via a charge transfer complex. *J. Biol. Chem.* **281**, 26847–26853 (2006).
14. Heyes, D. J., Sakuma, M., de Visser, S. P. & Scrutton, N. S. Nuclear quantum tunneling in the light-activated enzyme protochlorophyllide oxidoreductase. *J. Biol. Chem.* **284**, 3762–3767 (2009).
15. Heyes, D. J., Sakuma, M. & Scrutton, N. S. Solvent-slaved protein motions accompany proton but not hydride tunneling in light-activated protochlorophyllide oxidoreductase. *Angew. Chem. Int. Ed.* **48**, 3850–3853 (2009).
16. Heyes, D. J., Levy, C., Sakuma, M., Robertson, D. L. & Scrutton, N. S. A twin-track approach has optimized proton and hydride transfer by dynamically coupled tunneling during the evolution of protochlorophyllide oxidoreductase. *J. Biol. Chem.* **286**, 11849–11854 (2011).
17. Hoeven, R., Hardman, S. J. O., Heyes, D. J. & Scrutton, N. S. Cross-species analysis of protein dynamics associated with hydride and proton transfer in the catalytic cycle of the light-driven enzyme protochlorophyllide oxidoreductase. *Biochemistry* **55**, 903–913 (2016).
18. Archipowa, N., Kutta, R. J., Heyes, D. J. & Scrutton, N. S. Stepwise hydride transfer in a biological system: insights into the reaction mechanism of the light-dependent protochlorophyllide oxidoreductase. *Angew. Chem. Int. Ed.* **57**, 2682–2686 (2018).
19. Kavanagh, K. L., Jörnvall, H., Persson, B. & Oppermann, U. Medium- and short-chain dehydrogenase/reductase gene and protein families: the SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes. *Cell. Mol. Life Sci.* **65**, 3895–3906 (2008).
20. Korman, T. P., Tan, Y. H., Wong, J., Luo, R. & Tsai, S. C. Inhibition kinetics and emodin cocrystal structure of a type II polyketide ketoreductase. *Biochemistry* **47**, 1837–1847 (2008).
21. Javidpour, P. et al. The determinants of activity and specificity in actinorhodin type II polyketide ketoreductase. *Chem. Biol.* **20**, 1225–1234 (2013).
22. Blaise, M., Van Wyk, N., Banères-Roquet, F., Guérardel, Y. & Kremer, L. Binding of NADP<sup>+</sup> triggers an open-to-closed transition in a mycobacterial FabG  $\beta$ -ketoacyl-ACP reductase. *Biochem. J.* **474**, 907–921 (2017).
23. Zhao, F. J. et al. Crystal structure and iterative saturation mutagenesis of ChKRED20 for expanded catalytic scope. *Appl. Microbiol. Biotechnol.* **101**, 8395–8404 (2017).
24. Menon, B. R. K., Hardman, S. J. O., Scrutton, N. S. & Heyes, D. J. Multiple active site residues are important for photochemical efficiency in the light-activated enzyme protochlorophyllide oxidoreductase (POR). *J. Photochem. Photobiol. B* **161**, 236–243 (2016).
25. Klement, H., Helfrich, M., Oster, U., Schoch, S. & Rüdiger, W. Pigment-free NADPH: protochlorophyllide oxidoreductase from *Avena sativa* L. *Eur. J. Biochem.* **265**, 862–874 (1999).
26. Menon, B. R. K., Walther, J. P., Scrutton, N. S. & Heyes, D. J. Cryogenic and laser photoexcitation studies identify multiple roles for active site residues in the light-driven enzyme protochlorophyllide oxidoreductase. *J. Biol. Chem.* **284**, 18160–18166 (2009).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

# Article

**Acknowledgements** The work was funded by the Engineering and Physical Sciences Research Council (fellowship to N.S.S., EP/J020192/1). We thank Diamond Light Source beamlines I03 & I04 (proposal numbers MX8997-35, MX12788-8, 19, 42 and 62) and Shanghai Synchrotron Radiation Facility beamlines BL17U1 and BL19U1 for assistance during data collection. Time-resolved visible absorption measurements were performed using instrumentation funded by BBSRC Alert14 Award BB/M011658/1. This work was also supported by MOST 973 Project Biological Nitrogen Fixation (2010CB126504), National Basic Research Program of China (2014CB910304) and National Natural Science Foundation of China (31230004 and 81572090). We thank R. Dixon and R. Read for early discussions and support, and C. E. Bauer and Y. Fujita for kindly providing the strain ZY5 for the production of Pchlide.

**Author contributions** D.J.H., A.Z., Q.C. and N.S.S. initiated and coordinated the project. S.Z., Q.C., A.Z., N.S.S., D.L. and D.J.H. designed experiments, analysed data and wrote the manuscript with contributions from other authors. S.Z., M.S., L.F., W.S., H.L., J.Y. and X.L. produced and crystallized the proteins. A.Z. and C.W.L. collected and processed diffraction

data and solved the structures. S.H. and L.O.J. performed the docking and molecular dynamics simulations. S.J.O.H. and D.J.H. performed time-resolved spectroscopy measurements. R.H. and M.S. assisted with protein purification and characterization of POR variants. X.Y., M.L. and Z.R. advised on all aspects. All authors discussed the results and commented on the manuscript.

**Competing interests** The authors declare no competing interests.

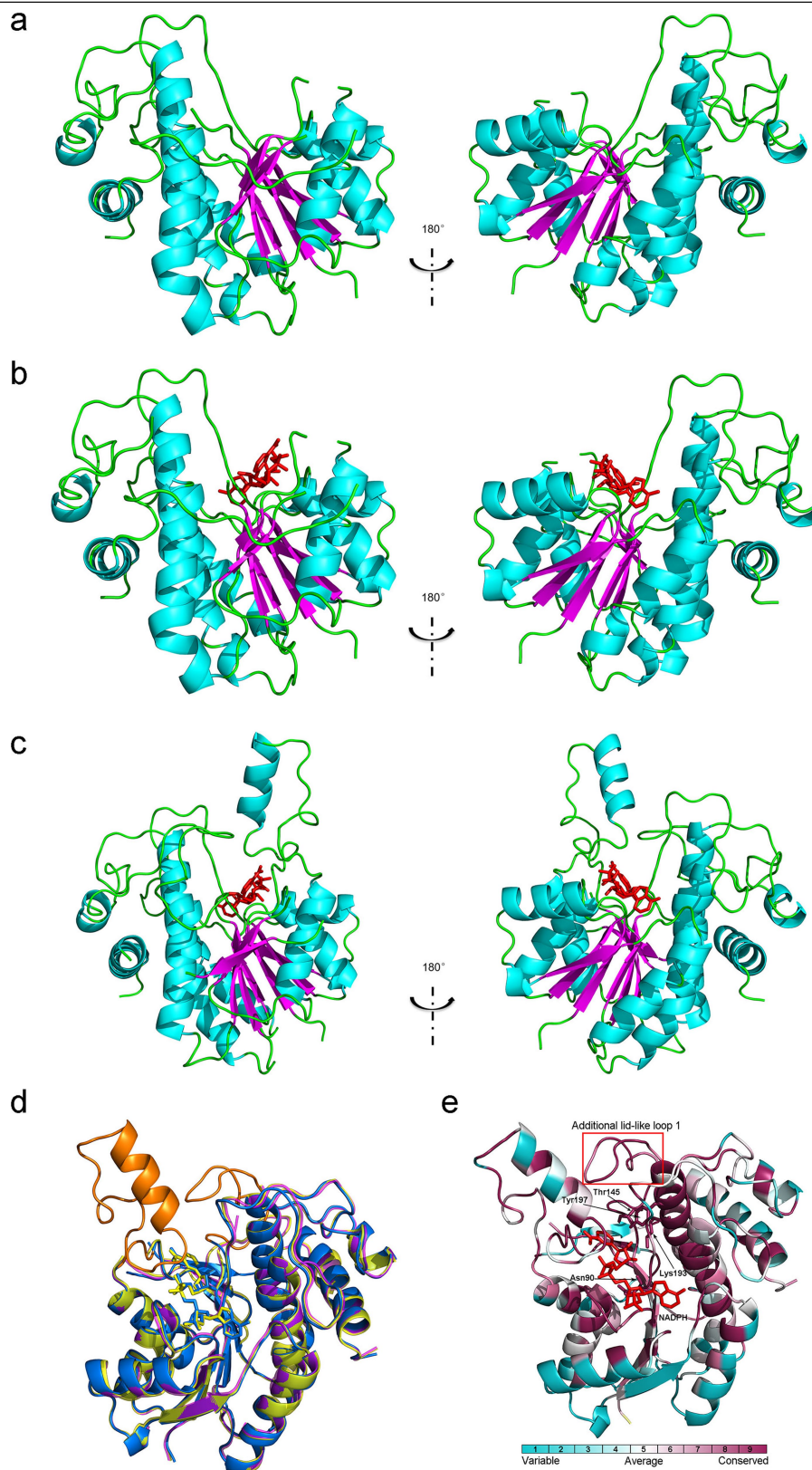
**Additional information**

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1685-2>.

**Correspondence and requests for materials** should be addressed to A.Z., Q.C. or N.S.S.

**Peer review information** *Nature* thanks Marco De Vivo, Sibongile Mafu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

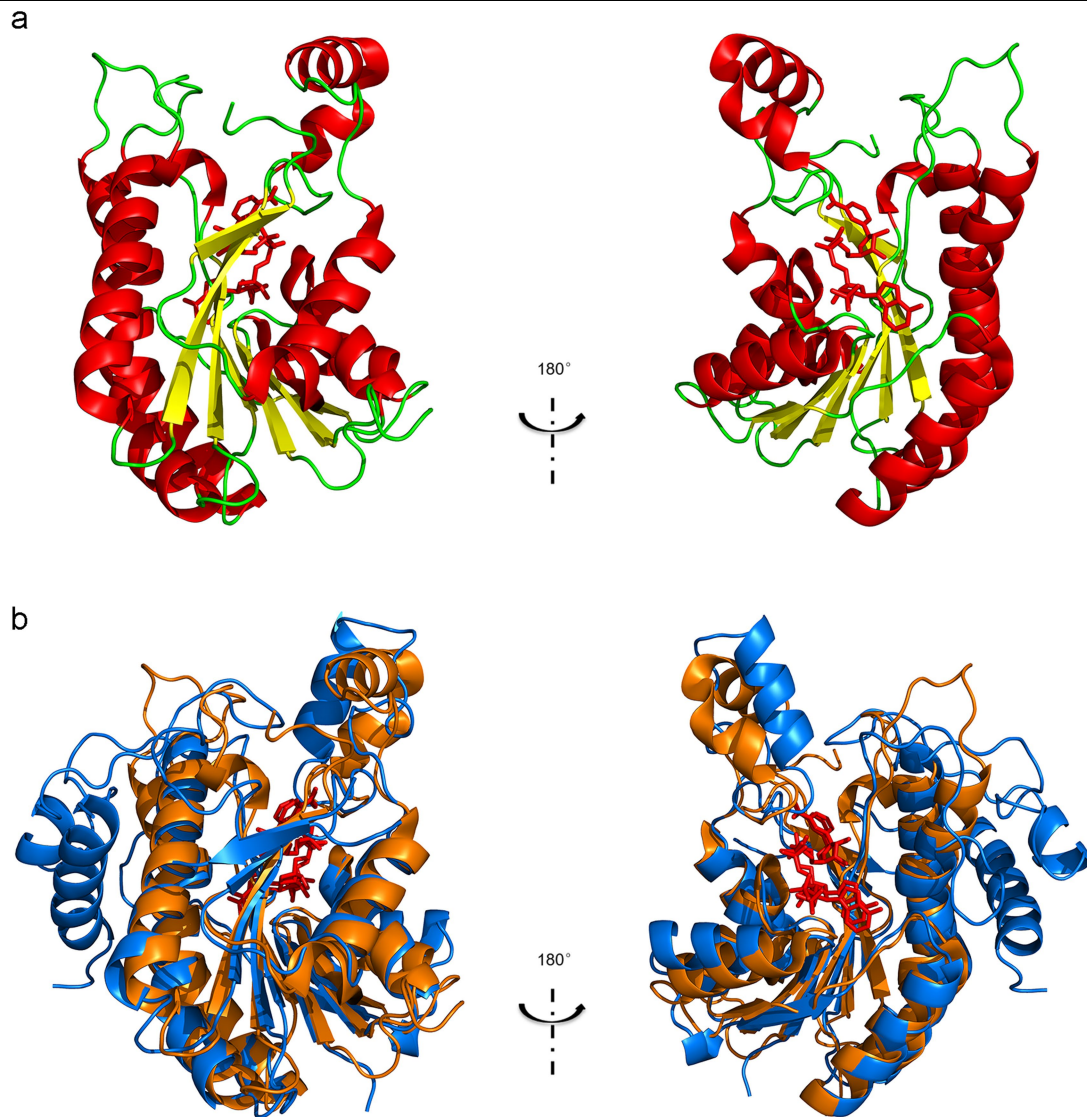
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.



**Extended Data Fig. 1 | Crystal structures of the *T. elongatus* and *Synechocystis* PORs.** **a**, *T. elongatus* apo-POR. **b**, *T. elongatus* POR with bound NADPH. **c**, *Synechocystis* POR with bound NADPH. The protein structure is coloured according to secondary structure; cyan,  $\alpha$ -helix; magenta,  $\beta$ -sheet; green, loop. **d**, Alignment of the POR crystal structures that we solve here. Magenta, apo structure of *T. elongatus* POR; yellow, structure of NADPH-bound

*T. elongatus* POR; blue, structure of NADPH-bound *Synechocystis* POR; orange, missing loops in *T. elongatus* POR (which are present in the crystal structure of the *Synechocystis* POR). **e**, Evolutionary conservation of *Synechocystis* POR structure. Evolutionary conservation of amino acid positions in the *Synechocystis* POR protein has been performed on the basis of the phylogenetic relations between homologous sequences, using the online ConSurf server<sup>27</sup>.

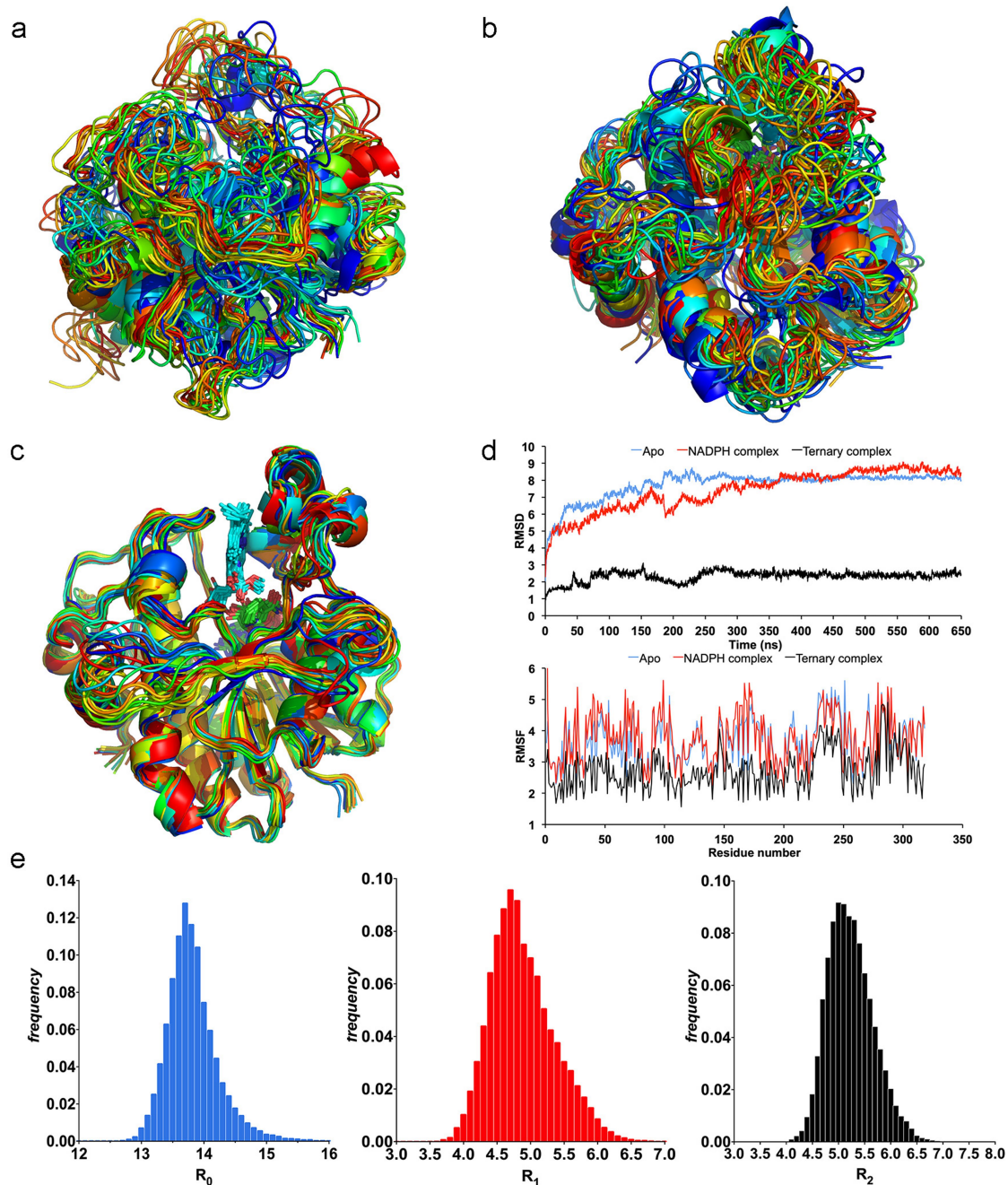




**Extended Data Fig. 2 | Comparison of the structure of the *Synechocystis* POR structure with that of the SDR-family protein  $\beta$ -ketoacyl reductase (PDB code 2B4Q). **a**, Overall structure of  $\beta$ -ketoacyl reductase protein (coloured by**

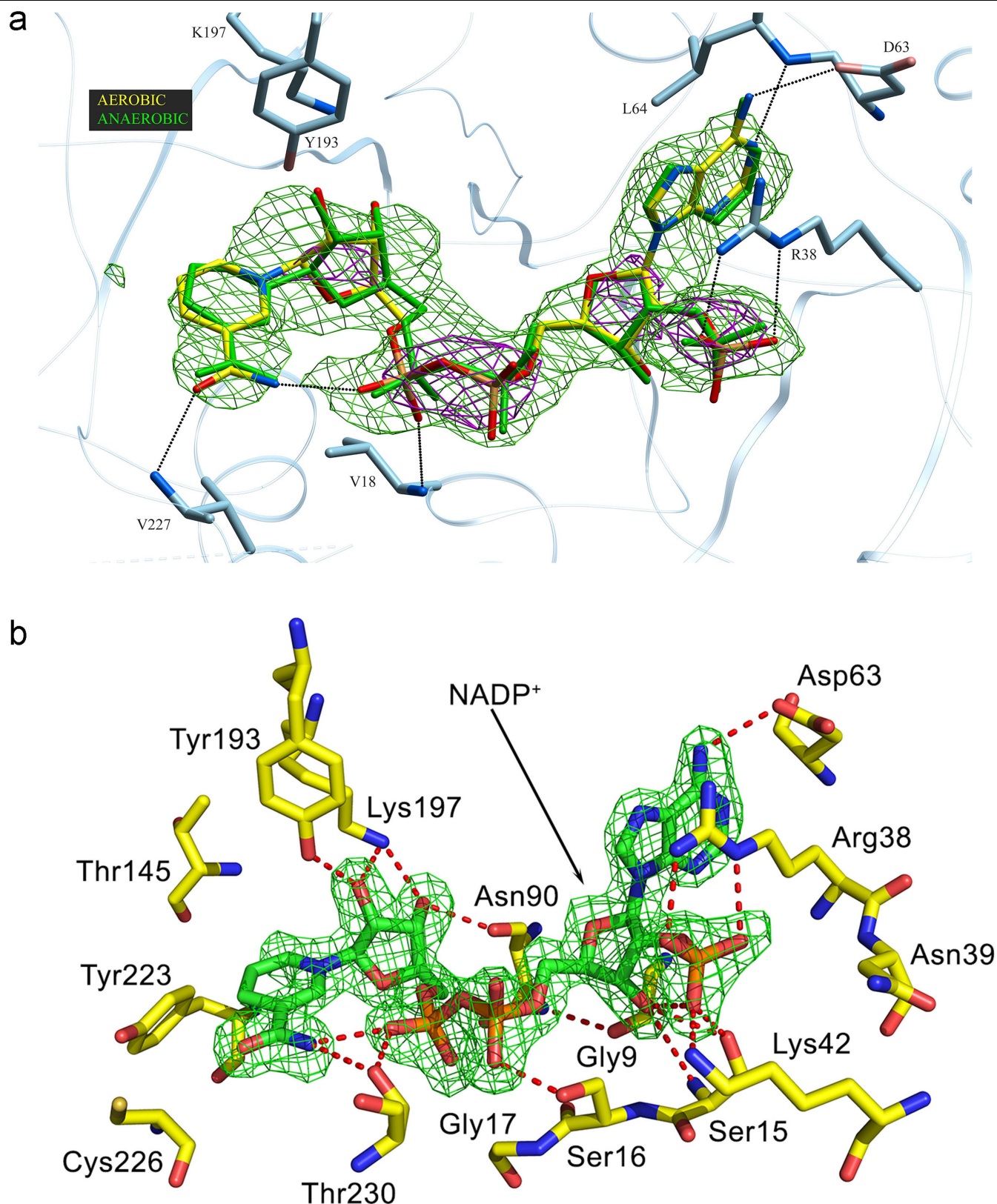
secondary structure). **b**, Alignment of the *Synechocystis* POR (blue) and  $\beta$ -ketoacyl reductase (orange). NADPH is shown as a red stick.





**Extended Data Fig. 3 | Molecular-dynamics simulations of the *T. elongatus* POR.** **a–c**, Overlays of structures from unrestrained molecular-dynamics simulations at 500-ps intervals for apo enzyme (**a**), POR-NADPH complex (**b**) and the ternary complex (**c**). **d**, Root-mean-squared deviation (RMSD) versus

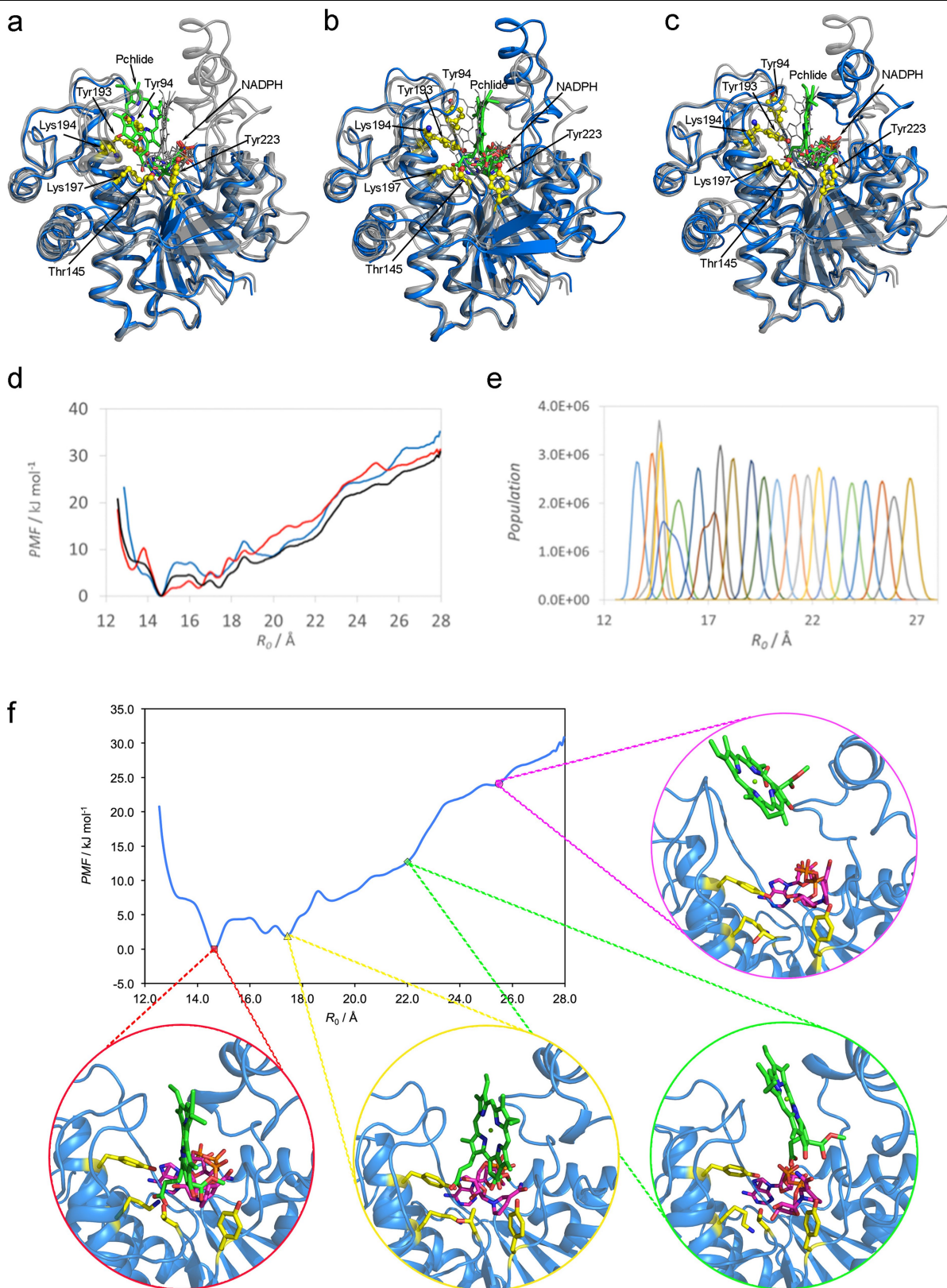
time (top) and the per-residue root-mean squared fluctuation (RMSF) (bottom), calculated for the non-hydrogen atoms of the protein. **e**, Distance distributions for the Pchlide-binding coordinate  $R_0$ , and the distances between the hydride donor and acceptor atoms ( $R_1$ ) and the proton donor and acceptor atoms ( $R_2$ ).



**Extended Data Fig. 4 | Electron density map for NADPH bound to the *T. elongatus* and *Synechocystis* PORs.** **a**, NADPH anaerobic soaking with the *T. elongatus* POR crystal. A low-resolution 3.5 Å electron density map (field-emission microscopy) contoured at  $1\sigma$  (green mesh) along with an  $F_o - F_c$  omit map contoured at  $4\sigma$  (magenta mesh) is shown for the NADPH region of an anaerobically soaked *T. elongatus* POR crystal. An all-atom coloured stick

representation of the aerobically soaked high-resolution NADPH-POR structure and associated active-site interactions are shown along with a stick representation of the anaerobically soaked NADPH in green. **b**, Structure of the NADPH-binding site of the *Synechocystis* POR. Hydrogen bonds between key residues and NADPH are shown as red dashes. The electron density for NADPH (omit  $F_o - F_c$  map contoured at  $3\sigma$ ) is coloured green.

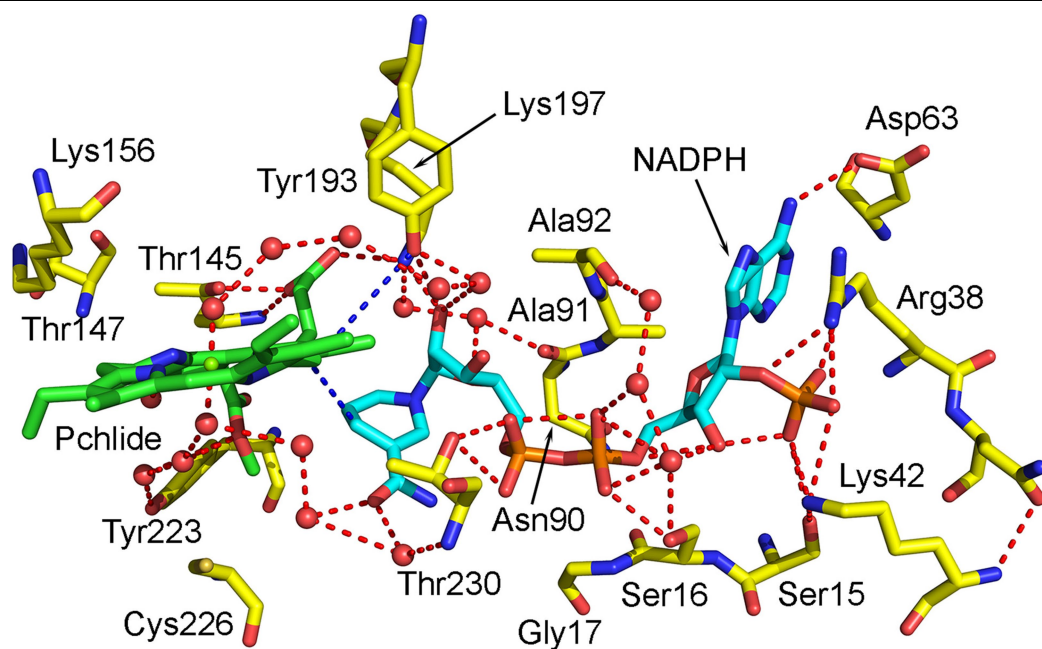




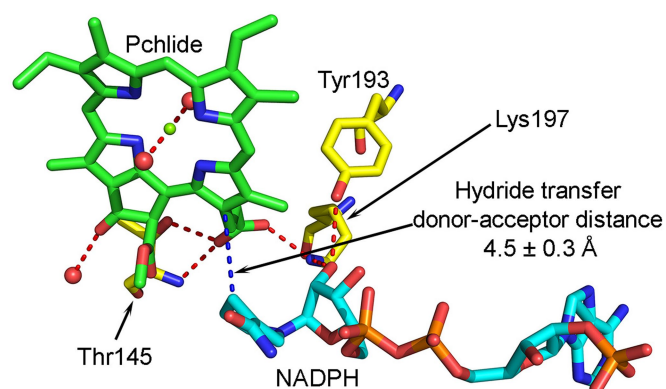
**Extended Data Fig. 5 | Modelling of the *T. elongatus* POR-Pchlide-NADPH ternary complex.** **a–c**, Three stages of modelling the *T. elongatus* POR-Pchlide-NADPH ternary complex. The structures are aligned and superimposed. Each panel highlights one of the three structures in blue (with the other two shown in grey): crystal structure with the chosen docked Pchlide structure (**a**); the structure after 20-ns annealing (**b**); and the final representative structure (**c**). The flexible residues during docking are shown as a yellow ball-and-stick representation, and the Pchlide molecule is shown in green stick representation.

**d**, Potential of mean force (PMF) calculated by umbrella sampling as a function of  $R_0$  (Supplementary Methods) for 50-ns umbrella sampling (black), as well as for the first 20 ns (blue) and the last 20 ns (red). **e**, Population distributions for each bin, each sampled for 50 ns. **f**, Potential of mean force was calculated by umbrella sampling as a function of  $R_0$  (distance between Pchlide  $Mg^{2+}$  and lower edge of the POR binding pocket) (Supplementary Methods). The dip in the potential of mean force at an  $R_0$  of about 17 Å corresponds to the formation of a hydrogen bond between Y223 and the Pchlide keto group.

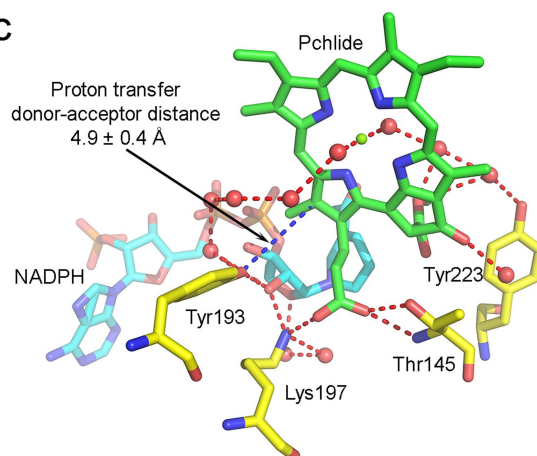
a



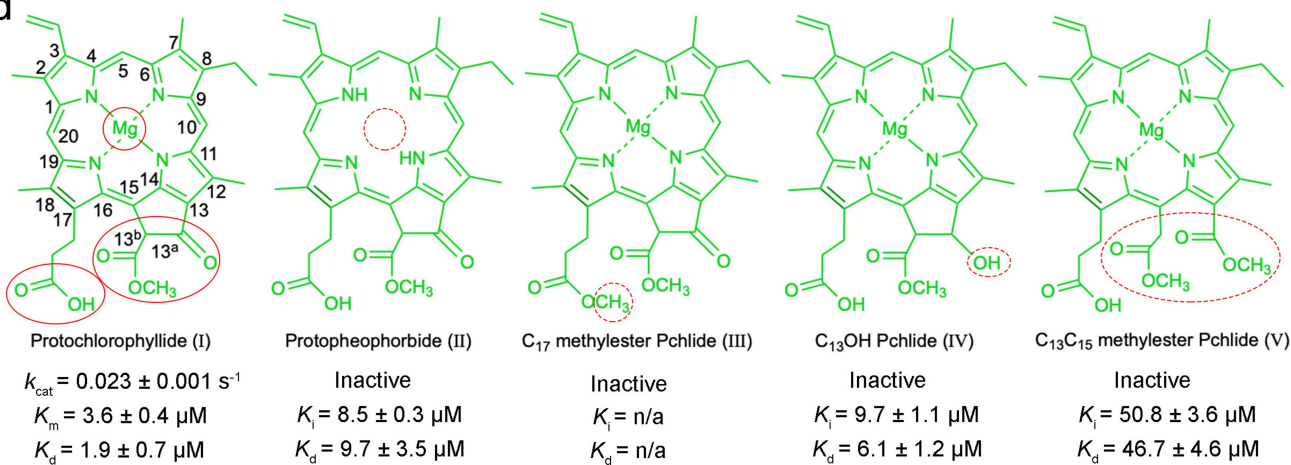
b



c



d



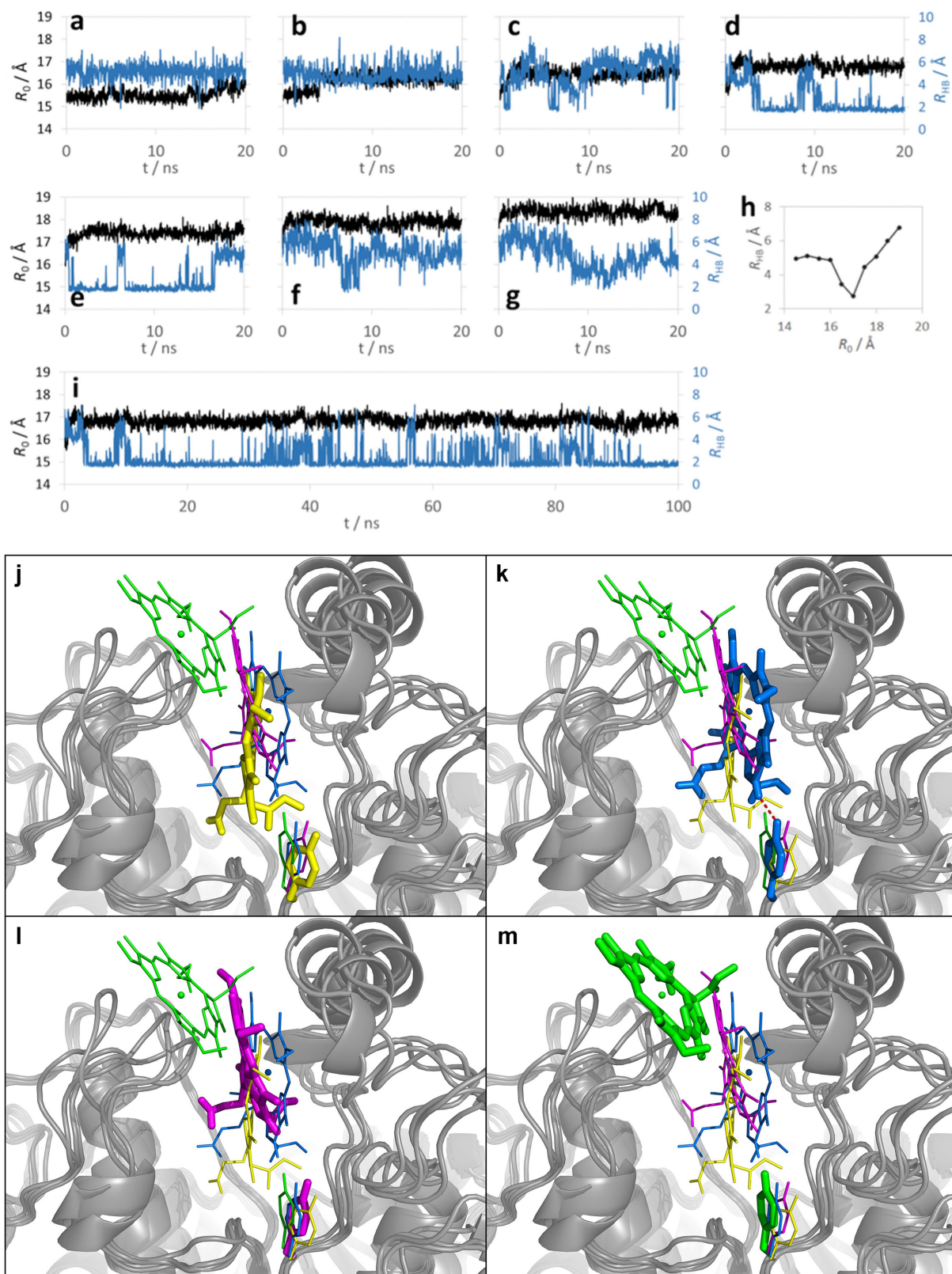
Extended Data Fig. 6 | See next page for caption.

**Extended Data Fig. 6 | Configuration of the active site and donor–acceptor distance of the *T. elongatus* POR–NADPH–Pchlde ternary complex model.**

**a**, Active site of the *T. elongatus* POR–NADPH–Pchlde ternary complex. The hydrogen-bonding network around the Pchlde and NADPH molecules is shown as red dashes. The donor–acceptor distance for hydride and proton transfer is shown as blue dashes. Water molecules are shown as red balls. **b**, View of the active site of the *T. elongatus* POR, highlighting the donor–acceptor distance for hydride transfer (shown as a blue dashed line). **c**, View of the active site of the *T. elongatus* POR, highlighting the donor–acceptor distance for proton transfer

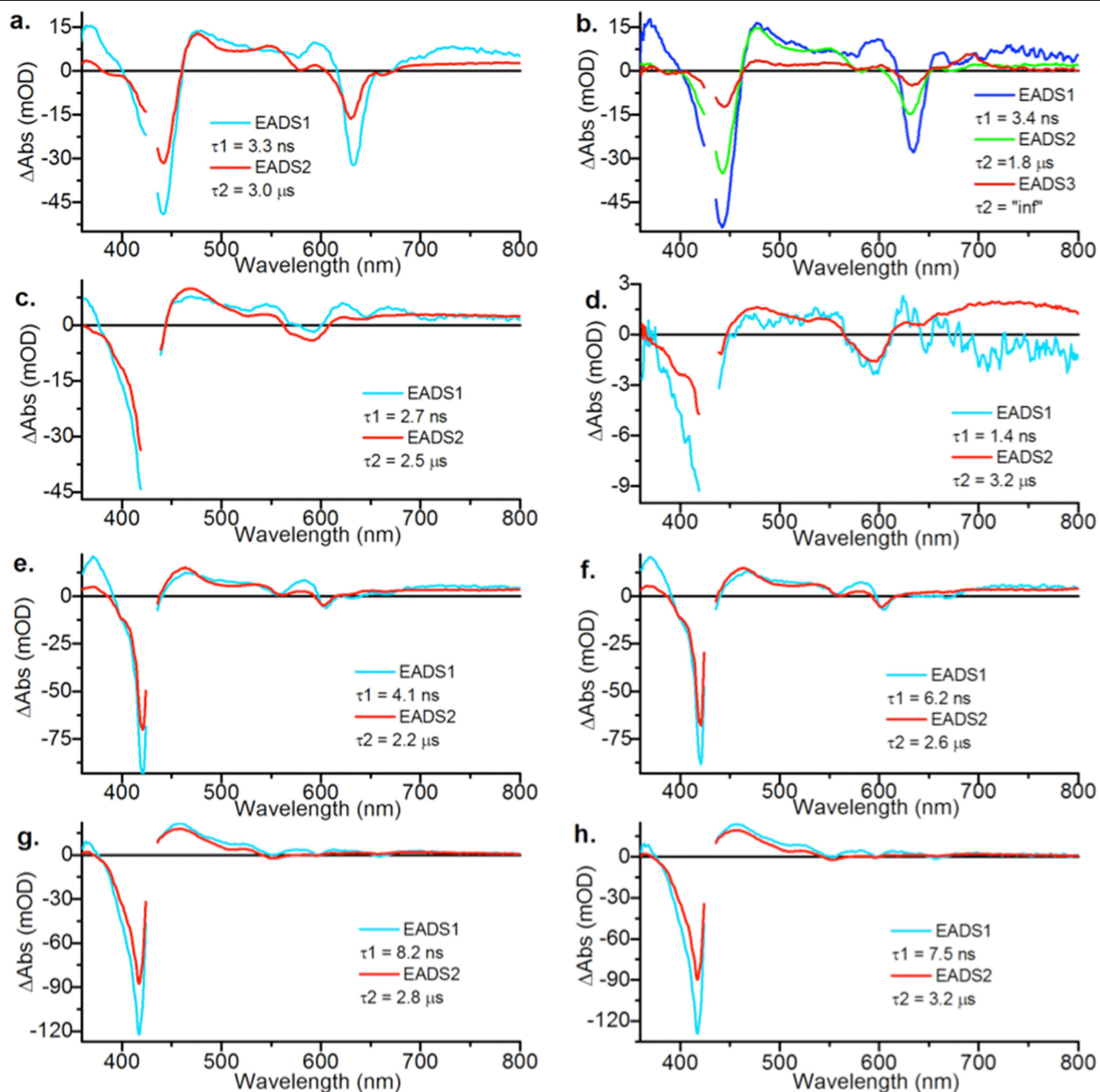
(shown as a blue dashed line). **d**, Summary of the activity, binding and inhibition data for the Pchlde analogues. The structures and apparent  $k_{\text{cat}}$ ,  $K_m$ ,  $K_d$  and  $K_i$  (where applicable, in each case) are shown for Pchlde (I), protopheophorbide (II), Pchlde with a C-17 methylester (III), Pchlde with a C-13–OH (IV) and Pchlde with a C-13 and C-15 methylester (V). The red circles show the regions of the Pchlde molecule that have previously been shown to be important for activity (central  $\text{Mg}^{2+}$ , ring E and the side chain at the C-17 position). The structures of all of the Pchlde derivatives described in the present study are shown with the modifications indicated by dashed red circles.





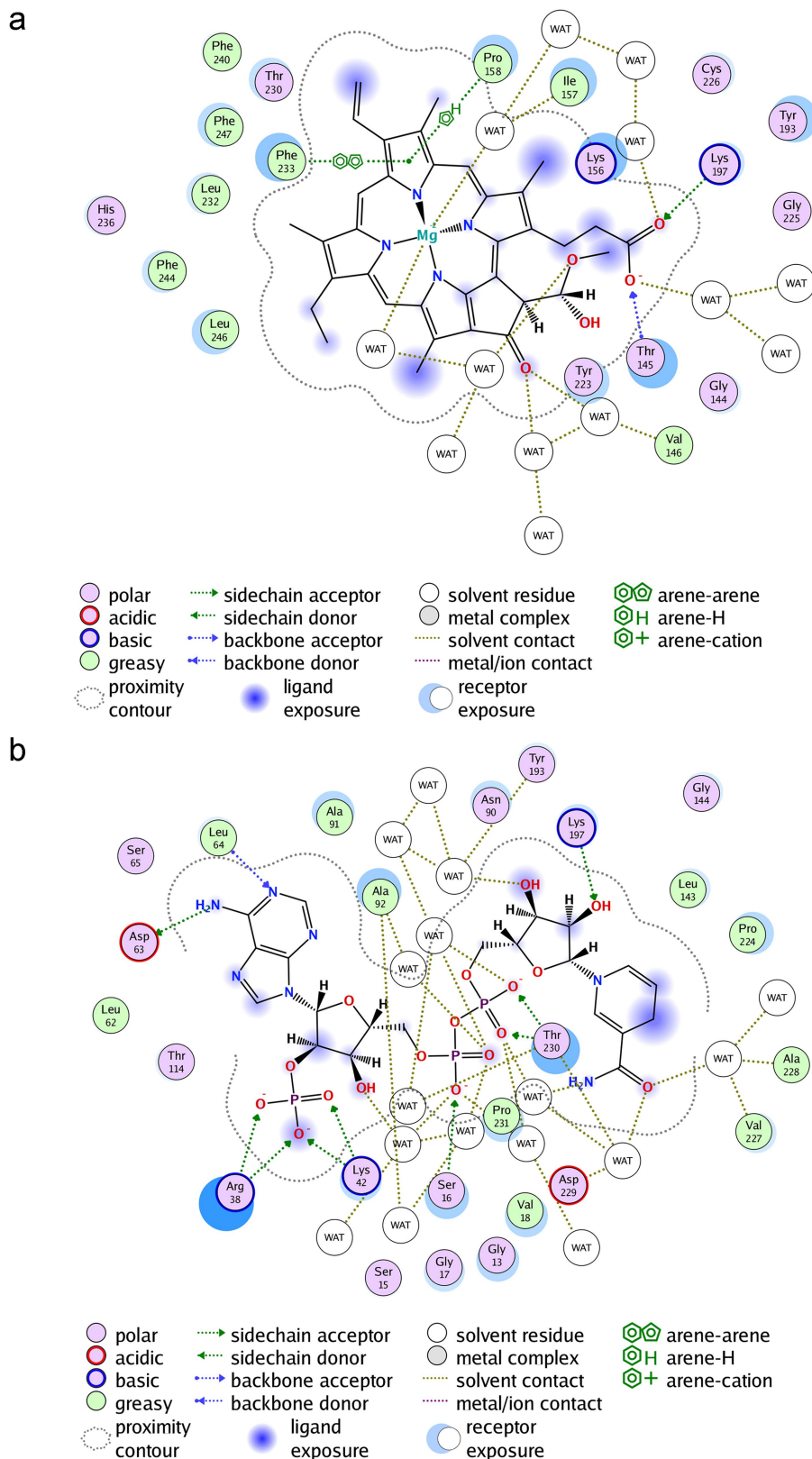
**Extended Data Fig. 7 | Hydrogen-bonding interactions between Tyr223 and the C-13 keto group during Pchlide binding.** **a–g,** Change in hydrogen bonding between Tyr223 and Pchlide C-13 keto group during Pchlide binding. Plots of the distance  $R_0$  between Pchlide  $Mg^{2+}$  and base of POR binding pocket (black), and the distance  $R_{HB}$  (HB, hydrogen bonding) between the Tyr223 hydroxy proton and keto oxygen (blue) from 20-ns molecular dynamics simulations at increasing  $R_0$  values. **h,** Average  $R_{HB}$  for each 0.5 Å bin for  $R_0$ . **i,** Extended 100-ns

molecular-dynamics simulation with  $R_0$  restrained at 1.7 Å to further illustrate the stability of the hydrogen bond between Y223 and the keto group. **j–m,** As the Pchlide leaves the binding pocket (shown sequentially from **j** to **m**), the Tyr 223 residue is free to move around and form a transient hydrogen bond with the C-13 keto group of Pchlide to 'guide' the Pchlide into its final orientation. The protein backbone is shown as a grey cartoon. At each stage, the Pchlide and Y223 molecules have been highlighted with thicker sticks in the figure.



**Extended Data Fig. 8 | Evolution associated difference spectra that result from global analysis using a sequential model of visible transient absorption data collected between 0.6 ns and 2.7  $\mu$ s. a–h,** Evolution associated difference spectra (EADS) are shown for Pchl<sub>a</sub> (a), POR with NADPH and Pchl<sub>a</sub> (b), protopheophorbide (c), POR with NADPH and protopheophorbide (d), Pchl<sub>a</sub> with a C-13 and C-15 methylester (e), POR with

NADPH, and Pchl<sub>a</sub> with a C-13 and C-15 methylester (f), Pchl<sub>a</sub> with a C-13–OH (g), POR with NADPH, and Pchl<sub>a</sub> with a C-13–OH (h). All data could be fitted using two EADS, except for POR with NADPH and Pchl<sub>a</sub>, which required three EADS owing to the formation of the hydride transfer intermediate (spectrum in red in b). The absence of any additional intermediates for the Pchl<sub>a</sub> analogues in the presence of POR implies impaired photochemistry.



**Extended Data Fig. 9 | Potential interactions in the *T. elongatus* POR ternary complex model. a**, The Pchlide molecule was chosen as the ligand, and the POR protein and solvent were chosen as the receptor. The 2D interaction map was calculated through molecular operating environment software (Chemical

Computing Group). **b**, The NADPH molecule was chosen as the ligand, and the POR protein and solvent were chosen as the receptor. The 2D interaction map was calculated through molecular operating environment software (Chemical Computing Group).

Extended Data Table 1 | Data collection and kinetic parameters of POR

a

	TePOR*	TePOR-NADPH	SsPOR-NADPH
<b>Data collection</b>			
Space group	P 3 <sub>1</sub> 2 1	P 3 <sub>1</sub> 2 1	P 2 <sub>1</sub> 2 2 <sub>1</sub>
Cell dimensions a, b, c (Å)	66.85 66.85 133.01	66.99 66.99 131.97	56.96 72.89 155.99
α, β, γ (°)	90 90 120	90 90 120	90 90 90
Resolution (Å)	53.08 - 1.27 (1.315 - 1.27)**	43.57 - 2.1 (2.175 - 2.1)	28.57 - 1.87 (1.91- 1.87)
R <sub>merge</sub>	0.0928 (0.8745)	0.1092 (0.7126)	0.097 (0.881)
I / σI	14.73 (2.36)	12.60 (3.18)	11.80 (2.1)
Completeness (%)	98.76 (96.46)	99.96 (100.00)	94.90 (86.50)
Redundancy	17.5 (12.1)	9.2 (8.9)	5.6 (6.0)
<b>Refinement</b>			
Resolution (Å)	53.08 - 1.27 (1.315 - 1.27)	43.57 - 2.1 (2.175 - 2.1)	28.57 - 1.87 (1.91- 1.87)
No. reflections	91463 (9068)	20674 (2023)	51836 (3223)
R <sub>work</sub> / R <sub>free</sub>	0.1526 / 0.1858	0.1672 / 0.1997	0.170 / 0.179
No. atoms			
Protein	2309	2224	4761
Ligand/ion	2	48	97
Water	363	158	416
B-factors (Å <sup>2</sup> )			
Protein	19.09	37.01	23.2
Ligand/ion	21.17	50.22	15.2
Water	34.55	41.16	32.4
R.m.s. deviations			
Bond lengths (Å)	0.010	0.007	0.019
Bond angles (°)	1.17	0.7	1.472

b

Pchlde and its analogue or TePOR variants	$k_{\text{cat}}$	$K_m/K_i$ Pchlde <sup>1</sup>	$K_d$ Pchlde	$k_{\text{hydride}}$	$k_{\text{proton}}$
	(s <sup>-1</sup> )	(μM)	(μM)	(×10 <sup>6</sup> s <sup>-1</sup> )	(×10 <sup>4</sup> s <sup>-1</sup> )
Pchlde (I)	0.023 ± 0.001	3.6 ± 0.4	1.9 ± 0.7	2.36 ± 0.11	2.23 ± 0.05
Protopheophorbide (II)	NA	8.5 ± 0.3	9.7 ± 3.5	NA	NA
C17 methylester Pchlde (III)	NA	NA	NA	NA	NA
C13OH Pchlde (IV)	NA	9.7 ± 1.1	6.1 ± 1.2	NA	NA
C13C15 methylester Pchlde (V)	NA	50.8 ± 3.6	46.7 ± 4.6	NA	NA
K156A	0.017 ± 0.001	3.11 ± 0.55	NA	2.40 ± 0.03	2.26 ± 0.14
Y223A	0.020 ± 0.004	33.2 ± 10.2	NA	1.77 ± 0.42	2.39 ± 0.46
Y223F	0.017 ± 0.001	11.6 ± 1.98	NA	1.80 ± 0.15	0.93 ± 0.07

**a.** Data collection and refinement statistics (molecular replacement). \*A single crystal was used for each data collection. \*\*Values in parentheses are for the highest-resolution shell. **b.** Kinetic parameters of the *T. elongatus* POR substrate analogues and variants. <sup>1</sup>K<sub>i</sub> values were measured for non-active substrate analogues. NA, not applicable for non-active substrate analogues. The values shown in the table are mean ± s.e.m., n = 3 for steady-state kinetics measurements, n = 5 for hydride and proton transfer measurements.



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Confirmed  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                                       |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated  |

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection: Generic Data Acquisition (GDA), MassLynx2.1

Data analysis: Origin 9.0, Xia2, Phenix 1.13, COOT 0.8.9.1, open-source PyMOL 1.8.7, Glotaran 1.5.1, Scwrl4, AutoDock Vina, Gromacs 5.0.4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The atomic coordinates and experimental data (codes 6G08, 6RNV, 6RNW) have been deposited in the Protein Data Bank ([www.wwpdb.org](http://www.wwpdb.org)). All other data are available from the corresponding author on reasonable request.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences      ☐ Behavioural & social sciences      ☐ Ecological, evolutionary & environmental sciences



## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size number was determined empirically from past experience based on previous results obtained with these particular methods & approaches.
Data exclusions	No data were excluded from the analyses
Replication	Three replicates were taken for the measurements, all attempts at replication were successful.
Randomization	Our study is about the protein structure, randomization is not relevant.
Blinding	Blinding was not relevant to our study. Our study is not about the efficacy of the compounds so there is no need to blind.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

# Prostate cancer: A declining art

Brachytherapy is an established treatment for prostate cancer with much to recommend it, but it is losing ground to flashier therapies. **By Michael Eisenstein**

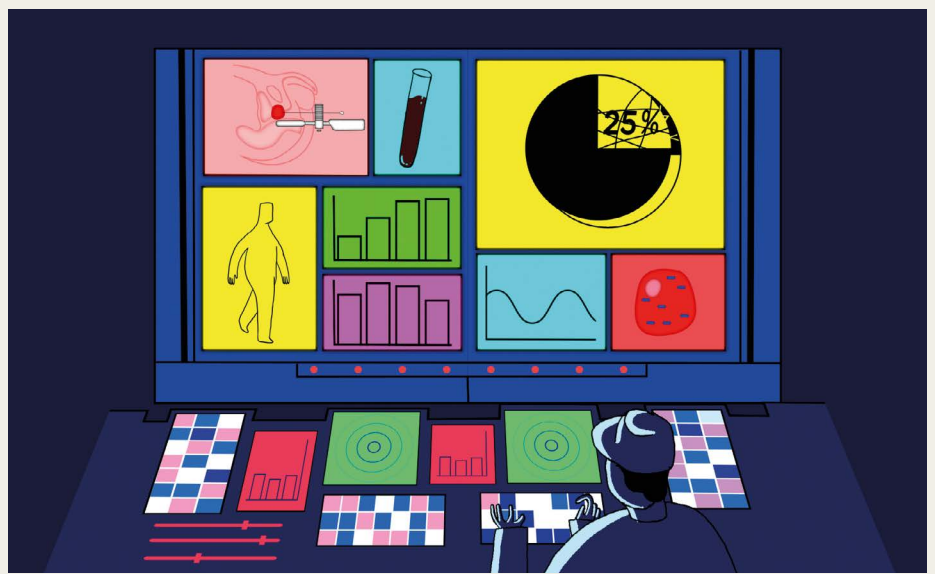
**M**odern medicine advances so quickly that it might be surprising to learn that a 100-year-old treatment for prostate cancer is still relevant. Brachytherapy, which involves bombarding the tumour with radiation from isotopes positioned around it, has a track record as a safe and effective procedure, and is less costly than other interventions such as robot-assisted surgery. Nevertheless, many oncologists are concerned that the technique could fall out of use.

In the first demonstration of brachytherapy in 1911, French physician Octave Pasteau used a urethral catheter containing radium. The technique evolved, and by the 1980s, a version known as low-dose-rate (LDR) brachytherapy, which is still in use today, had emerged. The therapy involves injecting 'seeds' of radioactive iodine or palladium into the gland, with the help of ultrasound imaging. These seeds are permanently embedded, releasing radiation for several months.

The process poses some problems: implanted seeds can expose patients' sexual partners to radiation, and the seeds can migrate into healthy tissues over time, for example. Another iteration of the treatment, known as high-dose-rate (HDR) brachytherapy, remedies this by temporarily introducing iridium isotopes into the prostate inside catheters.

LDR brachytherapy has long been used to treat people with prostate tumours, and the clinical performance of the HDR variety is promising. Both are delivered alone or alongside other treatments. But the use of both forms is in decline. In 2002, 17% of people in the United States with prostate cancer received the treatment; by 2010, that number had fallen to just 8% (J. M. Martin *et al. Cancer* **120**, 2114–2121; 2014).

In part, this decline can be ascribed to the fact that aggressive treatment by any method



ANDREW KHOSRAVANI

has become less common – many clinicians now opt instead to keep a close eye on low-risk tumours. But brachytherapy is also being eclipsed by more technologically sophisticated treatments such as robot-assisted surgery and proton therapy – a shift partly facilitated by hospital-reimbursement policies that favour newer approaches. The fall has alarmed many oncologists and radiotherapists, with some suggesting that it could lead to a decline in cure rates.

Without action, brachytherapy's decline in use seems set to continue. The drop means fewer opportunities for medical students and residents to see the technique in action – a feedback loop that, according to a survey of US radiation-oncology residents, might already be affecting their familiarity with the technique (N. Nabavizadeh *et al. Int. J. Radiat. Oncol. Biol. Phys.* **94**, 228–234; 2016). But

the radiation-oncology community is taking steps to ensure that brachytherapy remains an option. The American Brachytherapy Society in Reston, Virginia, has embarked on an initiative to train 30 practitioners in the technique every year for the next decade. And the American Society for Radiation Oncology is lobbying the US Centers for Medicare and Medicaid Services to re-evaluate how it reimburses hospitals for certain treatments. Brachytherapy has a long history, and many practitioners think that if it disappears, it is patients who will lose out.

*Nature* is pleased to acknowledge the financial support of the J-POPS Study Group, Nihon Medi-Physics Co., Ltd, the National Hospital Organization Tokyo Medical Center and the Translational Research Center for Medical Innovation. As always, *Nature* retains sole responsibility for all editorial content.

Produced with support from:



*Nature Outlines* are supplements to *Nature*, supported by external funding. They use infographics and animation to make complex scientific concepts accessible. *Nature* has sole responsibility for all editorial content (see [go.nature.com/2NqAZ1d](https://go.nature.com/2NqAZ1d)). Copyright © 2019 Springer Nature Ltd. All rights reserved.

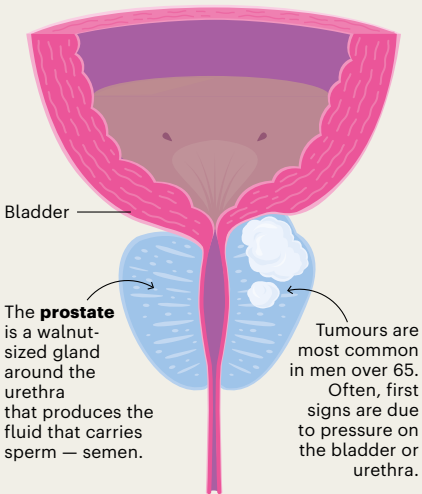
# KEEPING TREATMENT OPTIONS OPEN

People with prostate cancer currently have several treatment options available to them. But one of the oldest, brachytherapy, is losing popularity with physicians. Without action, the skills needed to perform this effective therapy could be lost.

By Michael Eisenstein; infographic by Lucy Reading-Ikkanda

## OLD THERAPY, NEW APPROACH

Brachytherapy uses radioactive material to kill cancerous cells. The conventional approach involves permanently implanting radioactive pellets in and around the prostate. In a newer form of brachytherapy, radioactive material is introduced temporarily through removable catheters. This might improve safety and efficacy, although only limited clinical data are available.



1 in 9 US men will develop prostate cancer.

Age-standardized cancer rate per 100,000 men<sup>1</sup>

31.5 (lung) 29.3 (prostate)

Globally in 2018, prostate cancer was the second most common cancer in men.

### Low-dose-rate brachytherapy

**Radioactive seeds** are injected into the prostate and surrounding tissue. The placement of the iodine-125 or palladium-103 seeds is permanent, and cannot be altered after injection.

**Dose:** Low-energy isotope seeds emit radiation over an extended period.

**Facilities:** Outpatient procedure with no specialized facilities.

**Administration:** Permanent implantation means greater care is needed with seed placement.

**Evidence:** Decades of clinical data support low-dose-rate (LDR) brachytherapy as a stand-alone and adjunct treatment.

**Risk potential:** Seeds can migrate to other tissues. Sexual partners might be exposed to radiation until the isotope decays.

### High-dose-rate brachytherapy

**Soft catheters** are inserted into affected areas of the prostate, and radioactive iridium-192 sources are placed in the catheters for each treatment session, which lasts around 2 hours. Multiple treatments are given over 1–2 days in hospital.

**Dose:** High-energy isotopes are applied in multiple, brief sessions.

**Facilities:** Specialized equipment and a shielded, radiation-safe site.

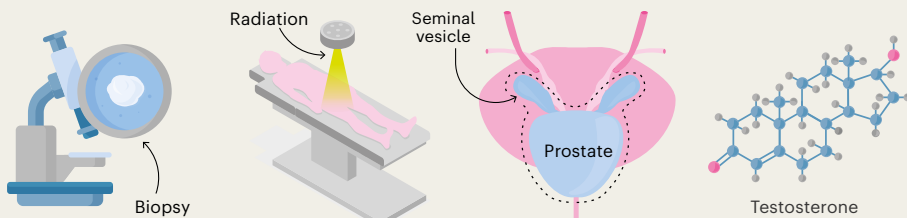
**Administration:** Isotope positioning can be adjusted between treatment sessions.

**Evidence:** Robust comparisons of high-dose-rate (HDR) brachytherapy with other therapies are still needed.

**Risk potential:** No radioactive material is left in the body after treatment, reducing the risk of problems later.

## A QUESTION OF RISK

Brachytherapy is one of several treatment options for prostate cancer. Treatment suitability depends on a tumour's level of risk, determined on the basis of factors such as size and cellular structure. Brachytherapy can be used alone for lower-risk tumours, or to boost more aggressive radiotherapy or hormone therapy in higher-risk cases.



### Active surveillance

Many clinicians now opt for careful, regular monitoring, rather than immediate treatment, for people with low-risk, slow-growing tumours.

### Beam therapy

For higher-risk disease, people are typically offered a more aggressive radiotherapy course, such as external-beam radiation therapy (EBRT).

### Radical prostatectomy

The prostate and seminal vesicles, which also secrete fluid found in semen, can be surgically removed to excise the tumour. This type of surgery is often robot-assisted for greater precision.

### Hormone therapy

Tumour growth can be stalled by reducing levels of hormones such as testosterone through drugs or surgical castration.

	Tumour risk		
	Low	Intermediate	High
Active surveillance	◆	◆	
EBRT	◆	◆	
Brachytherapy	◆	◆	
Radical prostatectomy		◆	◆
EBRT with brachytherapy		◆	◆
EBRT with hormone therapy		◆	◆
EBRT with brachytherapy and hormone therapy		◆	◆



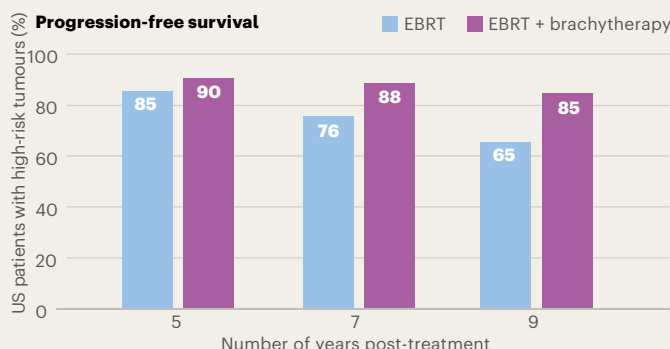
Watch an animation at  
[go.nature.com/2vkgz7j](https://go.nature.com/2vkgz7j)

## A VALUABLE OPTION

Randomized controlled trials have shown that brachytherapy has similar benefits as those of other treatments, but can be delivered at a lower cost and, in many cases, with more easily managed side effects.

### Survival gains

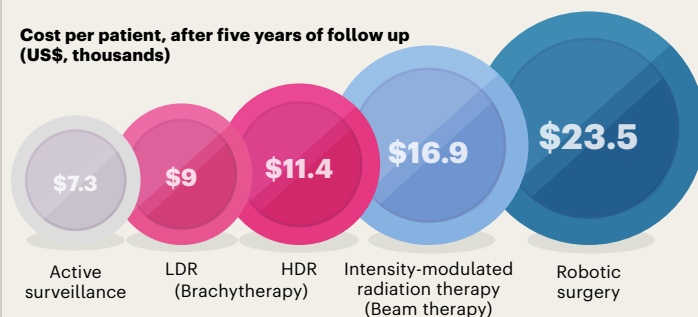
In people with low-risk tumours, brachytherapy is as effective (96%) as robotic surgery (97%) at preventing cancer recurrence within two years<sup>2</sup>. For high-risk disease, combining brachytherapy with EBRT boosts progression-free survival compared with EBRT alone<sup>3</sup>.



### Cost-effective

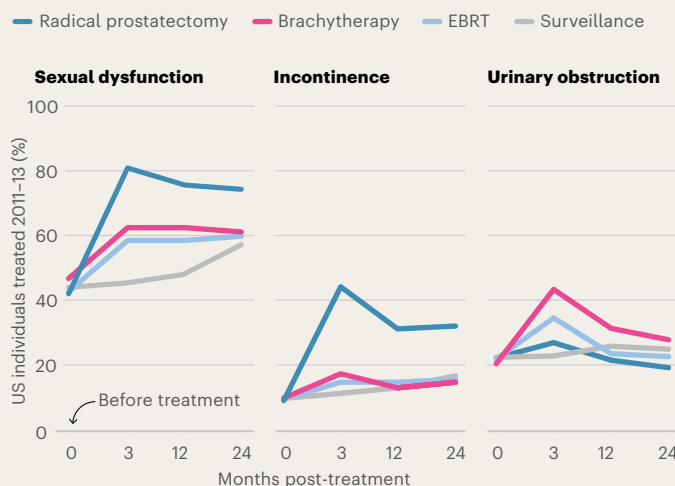
The cost of brachytherapy in US hospitals compares favourably with that of cutting-edge treatments such as advanced forms of beam therapy and robot-assisted surgery<sup>4</sup>. In Europe, treatment costs are more closely matched.

#### Cost per patient, after five years of follow up (US\$, thousands)



### Side-effect selection

Compared with surgery, brachytherapy tends to carry a lower risk of causing sexual dysfunction and incontinence. But it is associated with greater rates of urinary-tract obstruction than are other treatments<sup>5</sup>.

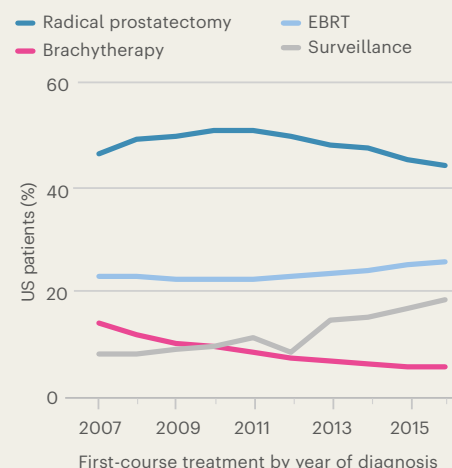


## SLIDE AWAY

Despite its apparent benefits, brachytherapy is struggling to compete with a growing clinical armamentarium. Worldwide, dwindling training opportunities put the future of this technique in jeopardy.

### Decreasing use

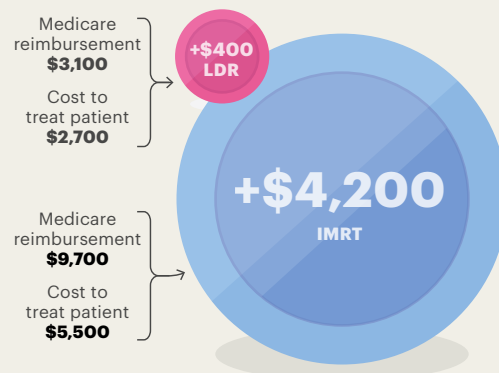
The proportion of US medical facilities performing brachytherapy has been falling each year since 2007, partly because clinicians are delaying treatment for low-risk malignancies in favour of active surveillance. But brachytherapy's use is also sliding compared EBRT and surgery<sup>6</sup>.



### Funding issues

In the United States, reimbursement rates by health insurer Medicare for brachytherapy (profit for LDR shown) are closer to the true cost than for other, more costly therapies, such as intensity-modulated radiation therapy (IMRT). This means the other therapies are more profitable<sup>7</sup>. In Europe, the reimbursement received by some hospitals doesn't always cover the cost of brachytherapy.

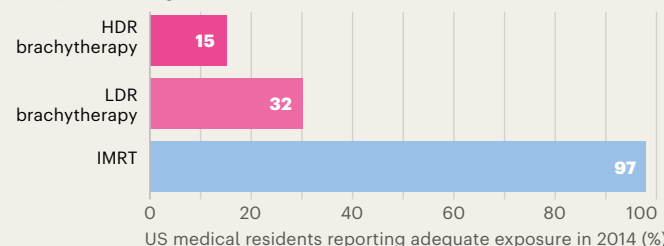
#### Profit per patient to hospital (US\$)



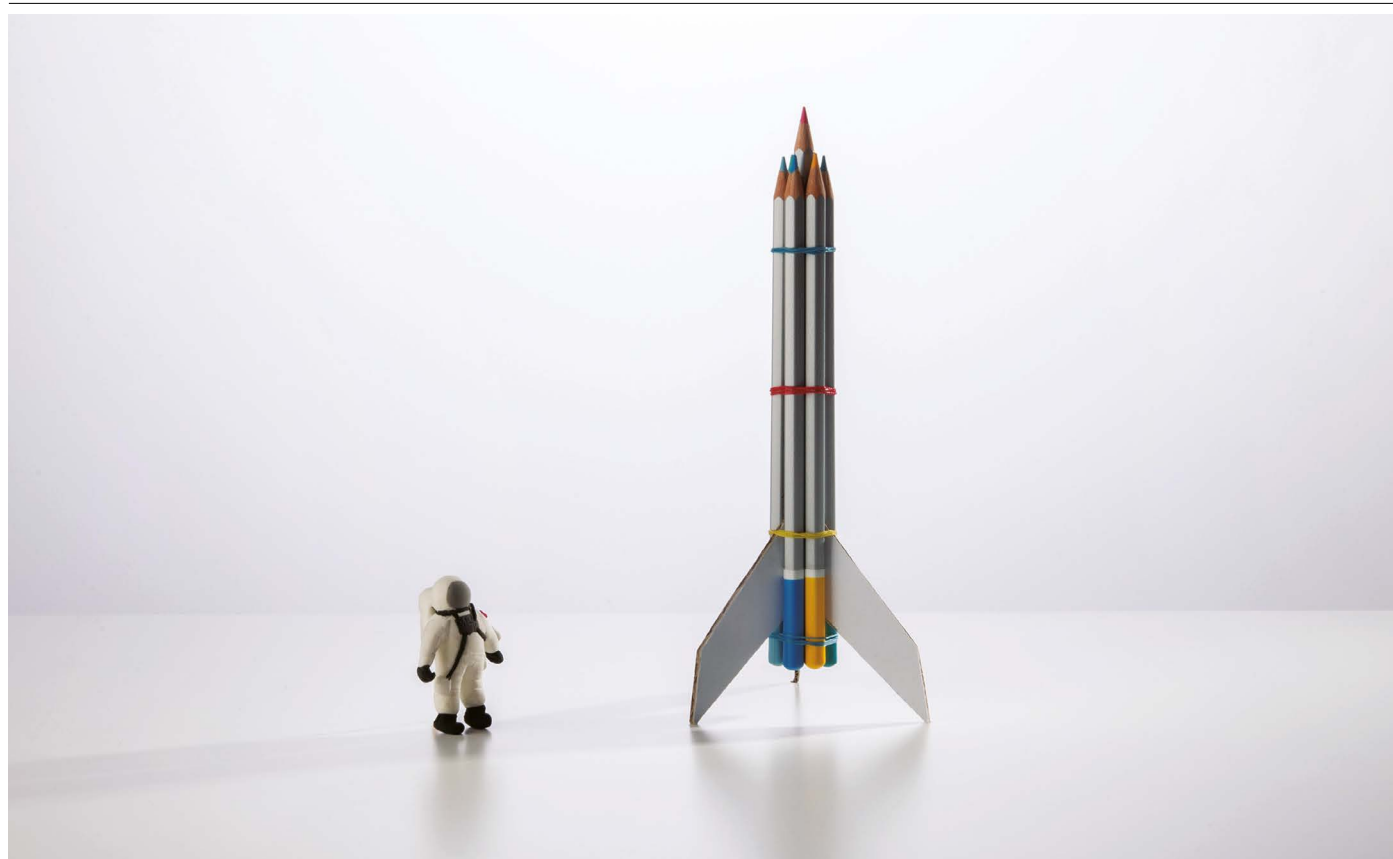
### Opportunity cost

As brachytherapy's popularity declines, so do the opportunities for clinicians early in their career to observe and learn the technique. US medical residents increasingly report insufficient opportunity for training in brachytherapy relative to newer therapies<sup>8</sup>. This could lead to further reductions in access.

### Adequate training in technique



SOURCES 1. GLOBOCAN 2018. 2. G. Claudio *et al.* *Can. J. Urol.* **24**, 8728–8733 (2017). 3. W. J. Morris *et al.* *Int. J. Rad. Oncol.* **98**, 275–285 (2017). 4. A. A. Laviana *et al.* *Cancer* **122**, 447–455 (2016). 5. R. C. Chen *et al.* *J. Am. Med. Assoc.* **317**, 1141–1150 (2017). 6. National Cancer Data Base. 7. S. W. Dutta *et al.* *Brachytherapy* **17**, 556–563 (2018). 8. N. Nabavizadeh *et al.* *Int. J. Radiat. Oncol. Biol. Phys.* **94**, 228–234 (2016).



SINO IMAGES/GETTY

## TAKING RESEARCH FROM DESIGN TO LIFT-OFF

If you want an astronaut to run your experiment in outer space, you have to keep it simple. **By Brian Owens**

**O**n 25 July, after years of planning, an experiment that Charles Cockell had spent years planning blasted into space. A SpaceX rocket launched from Florida, heading to the International Space Station (ISS). It carried 18 bioreactors, each the size of a deck of cards, that would be used to study whether bacteria could mine useful minerals on the Moon, Mars or asteroids.

Getting the experiment off the ground (literally) was “one of the most exciting things I’ve experienced”, says Cockell, an astrobiologist

at the University of Edinburgh, UK. But the process of getting from proposal to lift-off was long and involved: Cockell’s biomining experiment was more than 11 years in the making.

There are various routes to the ISS, but most go through one of the five space agencies – those in Canada, Japan, Russia and the United States, and the 11-nation European Space Agency (ESA) – that support the space station. These agencies periodically release calls for research proposals that help to meet their space-science goals.

“Most of our research has an eye toward enabling exploration, but it can also have terrestrial benefits,” says Craig Kundrot, director of NASA’s Division of Space Life and Physical Sciences Research and Application. The US portion of the ISS has been designated a national laboratory, meaning that it’s available for any research that would benefit from access to space, even if the project is not aimed at advancing space exploration.

Cockell’s experiment is intended to study how future Moon or Mars missions could use bacteria to extract materials from rocks in space, including minerals and metals for construction, water for rocket fuel, and soil.

Around 600 experiments are conducted each year on the station, which has extremely limited space, power and time allocated to astronauts to work on experiments. The payload specialists for research on the ISS – those who plan what to spend and when – have to juggle all of these factors as they decide which experiments can go when, and how these will all fit together. “They’re playing tetris on a daily basis,” says Kundrot.

Getting chosen by a space agency is just the start of the process. Then comes the tricky business of designing an experiment that can be packed into a rocket, blasted to the





NASA/NICK HAGUE

**NASA astronauts Christina Koch and Andrew Morgan stow biological research samples in a science freezer on the International Space Station.**

station and conducted by an astronaut who has hundreds of other responsibilities each day. It must also survive re-entry into the atmosphere and recovery, sometimes from hard-to-reach places such as the middle of the Pacific Ocean. Luckily, there are teams of engineering contractors and payload specialists that take the lead on making each experiment function within the limits on mass, size and power consumption.

“As a scientist, you’re not responsible for getting it to work. There’s a whole team who support the technical aspects,” says Monica Driscoll, a molecular biologist at Rutgers University in Piscataway, New Jersey. She was involved in a project running from December 2018 to January 2019 that used the nematode worm *Caenorhabditis elegans* to study the neurological effects of space flight.

Depending on the experiment, the engineering challenge can be relatively simple or fiendishly complicated. Physical-science endeavours tend to take longer to plan because they need bespoke equipment, whereas those in the biological sciences can often repurpose gear that worked in other projects. Driscoll had the benefit of drawing on the experience of previous work on *C. elegans* in space, but Cockell’s bioreactors had to be designed from scratch – part of the reason for the long timeline.

On the experimental design side, the challenge is to keep the project as simple as possible

while getting useful results. Something as basic as using a freezer, which wouldn’t require a second thought in a lab on Earth, can add another layer of complexity. The freezer on the ISS, which is designed to operate in microgravity, has a smaller storage capacity than a freezer in a domestic refrigerator, limiting the number of experiments that can use it. It has a waiting list,

**“If you want an astronaut to turn a knob, it has to be in their schedule.”**

says Kundrot. Plus, you have to work out how to keep the samples frozen during re-entry and recovery. That’s why Cockell’s team decided to forgo freezing its samples and instead opted to keep them in cooling packs for the return to Earth.

“Even simple things are complicated,” says Cockell. “If you want an astronaut to turn a knob, it has to be in their schedule.”

The experiment then undergoes several dry runs on Earth. One to verify that the science will work – for example, that Cockell’s bacteria were able to grow in the bioreactors – and then another, more detailed run-through of the operation of the experiment as if it was actually on the station. Once it has passed those tests to the satisfaction of the researchers, the space agency and the engineering contractors,

the experiment goes to a launch site, such as the Kennedy Space Center in Merritt Island, Florida, and is loaded onto a rocket. That, says Cockell, is when the excitement of what you are doing really hits home – as your work is readied to leave the planet.

“The station is not actually that far away – just 400 kilometres up, less than the distance from Edinburgh to London,” says Cockell. “But it’s the frontier of physical difficulty.”

Cockell’s experiment is now back on Earth, and he and his team are getting started on analysing the results. But they are already planning more work to send back to space. In two years, they will conduct another version of the experiment, this time bypassing the space agencies and using commercial contractors – Kayser Space in Didcot, UK, and its sister company Kayser Italia near Livorno, Italy. For £170,000 (US\$207,000), the researchers have bought access to the ISS for their experiments, without having to rely on the ESA peer reviewers to approve their work. Cockell says that it’s “just like buying a plane ticket to do research in another country”. Opportunities like this are opening up space research in a way that would not have been conceivable just a few years ago, he says. “It’s exciting for future generations. Within the next few years, this is something that a lot more people will be able to do.”

**Brian Owens** is a freelance science journalist based in St. Stephen, New Brunswick, Canada.





### Where I work Kerrie Mengersen

**A**s part of my work as a statistician, I research virtual habitats, including models of areas likely to be inhabited by jaguars. Here, I'm on Lake Imiria in Peru, in a wooden canoe made by Indigenous Shipibo villagers. Our team went out early in the morning to spot jaguar prey such as capybaras, peccaries and turtles, and to search for jaguars – or at least try to detect their calls or footprints.

It looked so serene with the huge trees and calm water, but the air was a cacophony of bird calls and mosquitoes – and was sometimes punctuated by shouts from team members when they glimpsed a sloth or a caiman. Boats are the best way to get around because the forest is very hot, very dense and potentially dangerous.

Because jaguars are rare and elusive, there are very few recorded observations of them. So my team at Queensland University in Brisbane, Australia, uses virtual reality (VR) to help understand them. We take photographs of selected sites that jaguars might inhabit and turn these photos into VR scenes. Then, instead of taking jaguar experts to the

jungle, we take the jungle to the experts. These are a mix of local Indigenous people, applying their knowledge of the region, and international experts. We immerse these specialists in different locations in our virtual jungle and ask them: "How likely is a jaguar to live, move through or hunt in this area?"

This immersive environment helps people to recall and identify important details that we need to build our statistical models. These predict where jaguars are most likely to roam, and are being used to guide conservationists in Peru who are building a corridor between protected areas.

For example, when the local people used our VR headset, they told us about the importance of specific fruit trees that the jaguars' prey rely on. I think of this human knowledge as data that are hiding in these experts' brains. The only way to tap those data is to put the experts right there, in the jungle scene.

**Kerrie Mengersen** is a statistician at Queensland University of Technology in Brisbane. **Interview by Kendall Powell.**

Photographed by  
Vanessa Hunter.